Journal of King Saud University – Computer and Information Sciences 31 (2019) 415–425

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Implications of big data analytics in developing healthcare frameworks – A review

Venketesh Palanisamy, Ramkumar Thirunavukarasu *

School of Information Technology and Engineering, VIT University, India

ARTICLE INFO

Article history:
Received 18 September 2017
Revised 20 November 2017
Accepted 7 December 2017
Available online 9 December 2017

Keywords:
Big data
Healthcare
Framework
Infrastructure
Analytics
Patterns
Tools

ABSTRACT

The domain of healthcare acquired its influence by the impact of big data since the data sources involved in the healthcare organizations are well-known for their volume, heterogeneous complexity and high dynamism. Though the role of big data analytical techniques, platforms, tools are realized among various domains, their impact on healthcare organization for implementing and delivering novel use-cases for potential healthcare applications shows promising research directions. In the context of big data, the success of healthcare applications solely depends on the underlying architecture and utilization of appropriate tools as evidenced in pioneering research attempts. Novel research works have been carried out for deriving application specific healthcare frameworks that offer diversified data analytical capabilities for handling sources of data ranging from electronic health records to medical images. In this paper, we have presented various analytical avenues that exist in the patient-centric healthcare system from the perspective of various stakeholders. We have also reviewed various big data frameworks with respect to underlying data sources, analytical capability and application areas. In addition, the implication of big data tools in developing healthcare eco system is also presented.

Contents

* Corresponding author.
E-mail address: ramkumar.thirunavukarasu@vit.ac.in (R. Thirunavukarasu).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2017.12.007
1319-1578/

1. Introduction

Big Data analytics and its implications received their own recognition in many verticals of which healthcare system emerges as one of the promising sectors (Andreu-Perez et al., 2015). The distinguishing characteristics of big data namely ***Volume*** (hugeness of data availability), ***Velocity*** (arrival of data as a flood of fashion), ***Variety*** (existence of data from multiple sources with diversified formats) find their own features in the abundant sources of healthcare data (Martin-Sanchez and Verspoor, 2014). The data sources for healthcare system have been broadly classified as (i) **Structured data**: Data that obeys defined data type, format, and structure. Example for such data in healthcare domain includes hierarchical terminologies of various diseases, their symptoms and diagnosis information, laboratory results, patient information such as admission histories, drug and billing information for the availed clinical services. (ii) **Semi-structured data**: Data that has been organized with minimal structure along with self-describing nature. Example for such data includes data generated from devices such as sensors for effective monitoring of patient's behaviour. (iii) **Unstructured data:** Data that has no inherent structure, which may include medical prescriptions written in human languages, clinical letters, biomedical literature, discharge summaries and so forth. Hence, the exploration of healthcare data to achieve valuable insights for diversified stakeholders (clinicians, patients, hospitals, pharmacy etc.) is a challenging and daunting task due to the enormous variety of data (structured, unstructured, semi-structured) from various sources. To extract ***Value*** (the fourth 'V' of big data attribute) from the existing 3 V's, it is appropriate to advocate efficient data processing platforms, smarter technologies for data collection, intelligent computational analysis, storage and visualization techniques (Sedig and Ola, 2014) towards attaining novel knowledge and efficient decision support strategies for different issues in healthcare.

Big data exhibits abundant potential to support a wide range of medical and healthcare functions such as clinical decision support, disease surveillance and population health management (Liang and Kelemen, 2016). Rapid advancement in Electronic Health Records (EHR) of patients, integration of social, behavioural and omics data with ICT based mHealth, eHealth, Smart Health and telehealth devices have led to the development of novel healthcare frameworks for supporting precision medicine and personalized patient care. Recent research attempts show that the comprehensive healthcare solutions have resulted with the aid of architectural frameworks that supports various levels of layered services. The underlying idea behind the development of these frameworks is the effective analytics of various healthcare data sources for identifying casual relationships and pattern of interest among the sources of big data. These frameworks promote healthcare solutions from the *disease centric model* to *patient-centric model*, where active participation of patients in their own healthcare resulted. In this paper, we review significant developments that occurred over recent years in building healthcare frameworks to orchestrate data analysis, management and visualization techniques as an end-to-end solution for the provision of improved healthcare services.

Due to the existence of diversified data formats, huge volume and associated uncertainty that exist among the sources of big data, the task of data curation plays a vital role in transforming raw data into an actionable knowledge. Though medical data are complex in nature, they exhibit strong inter-dependency and hence tasks such as simplification of data complexity, identification of interconnection among various health features, selection of target attributes for healthcare analytics require highly sophisticated and matured domain specific tools and techniques. There are various diversified complexities that exist in the healthcare computing environment such as handling of stream oriented data from ubiquitous devices for the purpose of patient monitoring, integration of disparate data sources towards developing predictive models, format construction and compression of medical images. In addition, most of the data enrichment activities and knowledge synthesizing processes in the various stages of data analytics life cycle largely rely upon the usage of big data tools. In this perspective, the important features of various big data tools that play significant role in the construction of healthcare frameworks are highlighted by providing a systematic review.

The organization of the paper has been structured as follows: The impact of big data analytics in healthcare system from the perspective of various stakeholders has been conceived in Section 2. Section 3 deliberately reviews various state-of-the art research attempts in developing healthcare frameworks along with their pros and cons. The implication of various big data tools in delivering healthcare solutions is highlighted in Section 4. Finally, Section 5 provides concluding remarks of the paper.

2. The impact of big data in healthcare

In the business sector, the core value of big data has been effectively utilized for the identification of behavioural patterns of the consumers to develop innovative business services and solutions. In the healthcare sector, the implication of big data serves predictive analytical techniques and machine learning platforms (Al-Jarrah et al., 2015) for the provision of sustainable solutions such as the implementation of treatment plans and personalized medical care. Jee and Kim (2013) compared the healthcare big data with the big data generated from the business sector under different attributes and their values. They redefined the characteristics of the healthcare big data into three features namely *Silo, Security, and Variety* instead of *Volume*, *Velocity* and *Variety*. Silo represents the legacy database that contains public healthcare information maintained in stakeholders' premises such as hospitals. The *security* feature implies the extra care needed in maintaining healthcare data. The *variety* feature indicates the existence of healthcare data in many forms such as structured, unstructured and semi-structured.

With the advent of big data analytics and its associated technologies, the healthcare domain witnessed pragmatic transformations at various stages from the perspective of involved stakeholders (Wang and Alexander, 2015). The impact of big data in healthcare results in identifying new data sources such as social media platforms, telematics, wearable devices etc. in addition to the analysis of legacy sources that includes patient medical history, diagnostic and clinical trials data, drug effectiveness index etc. When the mixture of these data sources and analytics are coupled together, it provides a valuable source of information for healthcare researchers towards attaining novel healthcare solutions (Zhou et al., 2017). A typical patient centric healthcare ecosystem with its significant stakeholders and their diversified data sources (structured/semi-structured/unstructured) is perceived in Fig. 1.

When these stakeholders work collaboratively and share their data insights effectively, healthcare solutions would be offered in

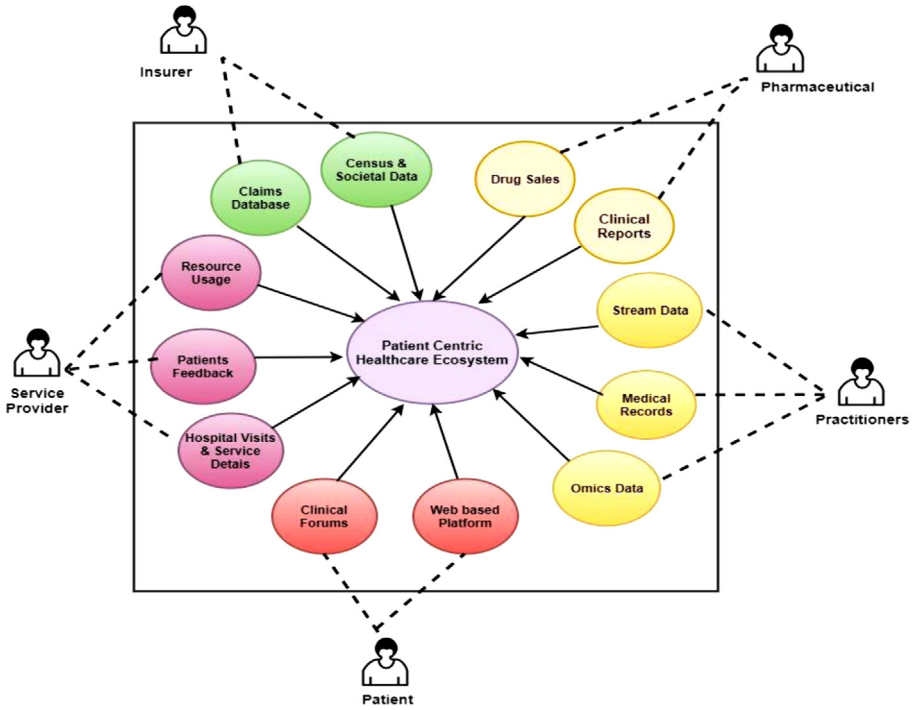

Fig. 1. A patient centric healthcare ecosystem – From the perspective of big data.

a cost-effective manner with improved personalized care for patients.

By considering the importance of these stakeholders in building big data healthcare ecosystem, the next section throws insights on their perspective over the effective utilization of big data sources.

2.1. Healthcare stakeholders & big data sources

This section elaborates how the diversified sources of big data cater the requirements of prominent stakeholders (patients, medical practitioners, hospital operators, pharma and clinical researchers, healthcare insurers) in attaining novel healthcare solutions. In addition, it offers appropriate analytical techniques to extract meaningful healthcare patterns of interest from the specified data sources.

2.1.1. Patients

The patient community always expects to avail a broad range of healthcare services at affordable cost with personalized recommendations (Mancini, 2014). In addition to physician's clinical diagnosis, they have an opportunity to gain more medical knowledge through digital platforms such as social media networks, clinical forums etc. These big data sources enable the patients to connect with similar people for gaining information such as disease symptoms, side-effects, hospitalization, drug information, feedback about clinical reports and post effect scenarios with improved privacy (Bouhriz and Chaoui, 2015; Kupwade Patil and Seshadri, 2014). Those patients who are unable to visit hospitals can avail telemedicine services for their healthcare needs. The platform could act as a big data repository and capture vital health signs such as temperature, heart rate, blood pressure and stream those data into a centralized repository for triggering periodical health alerts.

2.1.2. Medical practitioners

The massive amount of data generated from various phases of diagnosis and treatment plans of patients helps healthcare providers to identify the real insight about the progress of the treatments that are offered by them. There are many sources of big data that are generated by the healthcare system during the execution of treatment plans. It includes classification codes for various diseases and clinical services, laboratory results, clinical notes, medical imaging data, and sensor devices that capture patient's behaviour under different scenarios. When such big data sources are considered for constructing the Clinical Disease Repository (CDR), it improves public health surveillance and offers faster response through effective analysis of disease patterns. Besides, integration of data from wearable devices into healthcare applications also provides significant benefits such as facilitating physicians to track the usage of drugs, monitor the patient's health condition at any point in time.

2.1.3. Hospital operators

To effectively manage the patient's experiences and for optimizing the available resources, hospital operators highly rely upon the outcome of big data sources. Models based on predictive and prescriptive analytics are developed with the expertise of data scientists for measuring the strength of relationships between patient satisfaction indexes and availed services. Besides, resource allocation and optimization techniques can be successfully deployed on the basis of available big data towards fulfilling the manpower requirements for different sections of the hospital. The strategic operators of the hospitals can also utilize the location awareness data to decide the co-location of various departments to optimize the use of expensive healthcare equipment. Development of descriptive models on the basis of post-treatment data generated from follow-up phone calls, email communications, text messages would also facilitate to improve the offered services.

2.1.4. Pharma and clinical researchers

The impact of big data reflects a healthcare reformation in the domain of pharmaceutical and clinical research. The usage of omics and clinical big data (Wu et al., 2017) helps to build predictive models for understanding the biological and drug processes that

Table 1
Analytical avenues in a typical patient-centric healthcare system.

Stakeholders	Sources of big data	Nature of analytics	Underlying analytical techniques
		Classification of patient communities	Machine learning algorithms, statistical modeling techniques, sentiment analysis
Patients	Clinical forums and Telemedicine platforms	Health text analytics	Information retrieval techniques such as text mining
		Construction of health ontologies	Crawler based algorithms, semantic analytics
		Patient network analysis	Recommendation techniques based on collaborative filtering
	Genomic database	Genomic sequence analysis	Association rule mining and visualization techniques
Medical Practitioners	Electronic Health Record (EHR)	Categorization of patients based on personal statistics	Classification techniques based on machine learning algorithms
	Personal Health Record of patients	Identification of high-risk patient for specialized treatments, designing patient specific treatment plans, offering evidence based guidelines	Optimization and simulation techniques based on prescriptive analytics
	Device Generated data	On-demand real-time health analysis of patients	Analysing data streams by sampling and filtering techniques
Hospital operators	Electronic Health Record of patients, treatment plans	Prediction of inpatient duration of stay	Resource management and optimization techniques based on regression models and artificial neural networks
	Manpower utilization reports	Discovering strategies for resource planning and utilization	Optimization techniques such as evolutionary algorithms, PSO algorithms
	Patient's feedback	Computation of patient satisfaction scores based on demographical characteristics	Approaches based on user based and service based collaborative filtering
	Instrument usage log, calibration details	Detection and prediction of faults in medical devices.	Decision tree induction, Bayesian Classifier
Pharma and clinical researchers	Daily drug sales reports	Identification of usage and purchase patterns of drugs	Association rule mining, Time series analysis, Outlier mining
	Clinical reports, Patients Health Records	Amino acid sequence analysis, structure prediction of proteins for effective drug design	Deep learning techniques based on big data frameworks such as Hadoop and Spark
Healthcare insurers	Claim history	Trustworthy analysis of claims, demographic analysis of claims	Outlier analysis, Neural networks, genetic algorithms, nearest neighbor techniques
	Census and societal data	Proposal for new insurance plans	Algorithms based on predictive analytical techniques
	Data generated from Internet of Things	Reduction of claim frequency and severity	Techniques based on prescriptive analytics

attribute to the high success rate in attaining effective drug designs. Effective analysis of health data from diversified big data sources helps pharma companies to measure the outcome of designed drugs with smaller and shorter trials (Merelli et al., 2014). By coupling in-memory computing technologies with automated systems in drug manufacturing units, pharmaceutical companies can effectively integrate and analyze various forms of data to build end-to-end product solutions. Inputs from other stakeholders such as drug recommendation by a physician for a particular disease, quantum of consumption by patients, sales history from the drug shops enable the pharmaceutical organizations to evaluate and visualize their current market position for arriving strategic business decisions.

2.1.5. Healthcare insurers

The emergence of healthcare big data opens new analytical avenues for the benefit of healthcare insurers. Accordingly, novel health plans for frequently occurring diseases based on the geographical regions can be introduced with minimal premium cost. Advocating appropriate health plans for customers on the basis of various features such as age, gender, family history, income, nature of job enables benefits for both the insurer and customer. Analyzing unstructured data from claim history through predictive modeling techniques enables the insurance organization to predict patterns of authentic claims and unusual outliers for minimizing the cost of abuse. Big data coupled with Internet-of-Things (IoT) facilitate the insurers to introduce new and innovative business models such as *usage-based insurance* by analyzing the customer behaviour data captured in real-time. Mobile IoT (Internet of Things) plays a crucial role in transforming healthcare by allowing new business models to emerge and enables changes in work processes, productivity improvements and customer experiences(Dimitrov, 2016).

By analyzing the influence of big data on the above stakeholders, we have identified the potential big data sources in a typical patient-centric healthcare system. Table 1 presents various analytical strategies that explore meaningful patterns of interest in the domain of healthcare.

In the next section, we have reviewed various healthcare frameworks by highlighting their data sources, analytical capability and application areas.

3. Big data frameworks for healthcare

Recent research attempts advocate various healthcare frameworks for handling large volume of diversified data from disparate data sources to churn out significant patterns and trends. This section reviews those big data frameworks and highlights the contribution in the domain of healthcare.

An applied architectural framework for healthcare system using big data analytics has been proposed by Raghupathi and Raghupathi (2014). The framework consists of layers such as Data Source layer, Transformation Layer, Big Data platform layer and Analytical layer. The data source layer mainly focuses on internal and external data sources of healthcare found in multiple locations under various formats. The transformation layer is responsible for operations such as extraction, transformation, and loading of data into big data platform through various data staging techniques such as middleware and data warehousing operations. The layer of big data platform comprises of various Hadoop ecosystem tools

for performing specific operations on Hadoop Distributed File System (HDFS) using Map-Reduce programming model. The analytical layer performs operations such as querying, reporting, online analytical processing and data mining techniques. Besides, the authors have outlined various tools and platforms for analyzing healthcare big data. Though the proposed architectural framework is a pioneering one in the context of big data for the domain of healthcare, it only emphasizes the theoretical aspects. No experimental evaluations have been conceived based on the proposed framework.

Chawla and Davis (2013) proposed a patient centric personalized healthcare framework based on collaborative filtering approach. It captures patient similarities and produces personalized disease risk profiles for individuals. Collaborative filtering is a data analysis technique designed to predict user's opinion about an item or service based on the known preferences of a large group of users. In the proposed framework, individual patient's healthcare history has been compared with all other available patients' medical histories on the basis of defined similarity constraints such as occupation, symptom, lab result, family history and demographic data. Based on the similarity computation, a pool of similar patients is selected and prediction of diseases has been carried out. With the increasing use of electronic healthcare records, the proposed framework provides a proactive healthcare solution in the context of big data. Besides the advantage of offering patient centric personalized healthcare framework for physicians to assess the disease risk of patients, it handles only the diagnosis codes that confirm to ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) standards.

A big data analytical framework that utilizes ubiquitous healthcare system has been attempted by Kim et al. (2014). The framework analyses vital signs extracted from accelerometers to provide healthcare services. Vital signs are continuous time series data that are unstructured in nature and having inadequacy to store in the traditional databases. Electro Cardiogram signals (ECG), respiration and the motion data have been accounted as vital signs. The framework employed open standard platform to support interoperability among data and different devices. Hadoop platform has been extended by adding algorithms to extract feature values from raw data of vital signs and store them for real-time analysis. Despite the novelty of proposing Hadoop based platform for extracting and processing bio signals such as ECG, their work lacks in delivering substantial analytical models on the top of Map-Reduce programming model.

The implication of computational aspects in medical and health big data informatics have been extensively surveyed by Fang et al. (2016). They advocated a framework called 'Health informatics processing pipeline framework' that combines a sequence of steps to reap meaningful patterns from healthcare big data. The framework consists of process pipeline such as data capturing (identifying data sources such as electronic healthcare data, clinical support data sources, and laboratory results), storing (identifying cost effective storage infrastructure for analysing healthcare data), analysing (performing tasks such as data pre-processing, feature selection and machine learning) searching (extracting meaningful patterns of interest from the outcome of analysis), decision support (utilizing the pattern base for effective decision making in the health informatics domain). Besides the proposed framework, certain research directions with respect to issues pertaining to data heterogeneity such as structured and unstructured healthcare data, complexity existing in the available data, privacy issues and visualization of discovered patterns are also explored in their work. Their proposed healthcare framework offers a systematic data processing pipeline for stages of big health informatics such as data gathering, storing, searching and analyzing data from diversified sources. However the focus towards technological aspects of implementation with the aid of big data tools and techniques is obliterated.

With the advent of mobile devices and sensor networks, pervasive healthcare services emerged as a novel solution in the health informatics domain, since it offers healthcare services to the patient at anytime/anywhere basis. With the wide range of mobile devices, sensors and wearable applications, data generated in heterogenic formats have been utilized for the provision of on-demand healthcare services. A framework for healthcare big data analytics in mobile cloud computing environment was proposed by Youssef (2014). The framework provides high-level integration, interoperability, availability, and sharing of healthcare data among various stakeholders namely medical practitioners, patients, and drug developers through the following components: (i) Cloud component – hosts patient information and offers healthcare services (ii) EHR component – responsible for integrating distinct patient records from different sources such as pharmacy, hospital, and lab. (iii) Security component – guarantees the protection of security and privacy issues by implementing encryption and authentication techniques (iv) Data analytical component – deploys different analytical tools for discovering new kinds of patterns from the available HER (iv) Care Delivery Organization (CDO) components- represents the different healthcare organizations distributed in various locations. All such organizations can perform data sharing using HL7 protocol, which is the standard structure of communication among healthcare organizations. Though the framework elaborately discusses the security aspects for protecting patient data in a cloud environment, empirical evaluation of the proposed security policies have not been advocated.

The need for self-caring services for patients under emergency situation has been focused by Lin et al. (2015) and advocated a cloud based big data healthcare framework. It consists of an off-line Hadoop cluster and an online distributed search cluster. The Hadoop cluster is responsible for off-line storage and index building of medical documents and the online cluster has been designed for processing user query in a highly concurrent and scalable fashion. The online cluster consists of (i) Search node – for retrieving medical records (ii) Node for Data analysis – for developing disease symptom lattices (iii) Access control lattice – for filtering privacy information of patient (iv) Load balanced cluster – For balancing the load of user queries. Based on the proposed framework, a prototype design for home diagnosis service has been attributed for testing the validity of the proposal. Also, the scalability of the cloud based framework has been achieved by dynamically adding and removing the nodes in each cluster. Though the proposed work implements a prototype of cloud-based framework for self-diagnosis service, it analyzes only the historical medical records of patients.

The importance of semantic interoperability among clinical information is the prime focus of the work carried out by Legaz-Garca et al.(2016). They have argued that the lack of interoperability among clinical models and clinical record yield inefficiency in the healthcare system. To establish semantic integration of Electronic Healthcare Data (EHR), they proposed a framework based on Web Ontology Language (OWL). The patient data (EHR) obtained from the relational databases is transformed into OWL for ontology construction and the constructed ontology has been utilized for data exploration such as EHR based data classification and visualization. Though the framework offers advantages of using semantic technologies in biomedical research, adding further efforts in learning optimal set of parameters for constructing ontologies in archetypes will enhance the outcomes of framework.

A cyber-physical system based healthcare framework- 'smart healthcare framework' that integrates sensing technologies, cloud computing, Internet of things, and big data analytics has been proposed by Sakr and Elgammal (2016). Various layers that

constitute the framework are: (i) Data connection layer – for sensing, extraction and integration (ii) Data storage layer – for storing relational, non-relational and cloud oriented data (iii) Big Data processing and analytical layer – for performing various analytics such as descriptive, predictive and prescriptive analytics (iv) Presentation layer – for developing graphical dashboards and work flows. The proposed architecture holds good for various use-cases such as patient profile analytics, population management, genomic analytics and improved patient monitoring where integration of the above four technologies played a vital role. Their work significantly integrates various ICT advantages such as sensor technologies, cloud computing, Internet of Things and Big Data analytics in offering Smart Healthcare services. In other side, the framework lacks analytical capability to handle complex data sources such as images and streams.

A cloud based context aware framework to identify the impact of socio-economic, demographic and geographical conditions on public health has been attributed by Mahmud et al. (2016). It is an Amazon web service based cloud platform integrated with geographical information system for capturing, storing and visualizing the big data. Accordingly, contextual and healthcare data from various remote locations and regions have been captured and a predictive model based on fuzzy-rule based summarization technique for health-shock has been proposed. The fuzzy-rule based technique is used to generate interpretable linguistic rules for classifying the health shocks. The proposed cloud based model facilitates the healthcare professional to understand the impact of socio-economic, environmental and cultural norms that directly or indirectly caused the health-shocks. The novelty of the proposed work has been illustrated with real-time data sets collected from the rural and tribal areas of Pakistan. Though the paper advocates Fuzzy based predictive analytical framework, no adequate comparisons have been made with any benchmarking experimental studies.

Jokonya (2014) proposed an integrated big data framework that assists for the prevention and control of diseases such as HIV/AIDS, Tuberculosis, and silicosis. Their framework mainly focuses on the Mining Industry that creates an abundant amount of silica dust. The silica dust that affects the lungs may also cause silicosis and tuberculosis. When the immune system is compromised by HIV/AIDS and silicosis, it makes easier for the tuberculosis to infect the body. The proposed framework addresses the need of Epidemiology predictive model for forecasting and controlling the above mentioned diseases. The big data component of the framework performs data capturing of different Tuberculosis risk factors and combines the data for effective analysis through predictive and descriptive models. Though the proposed framework effectively performs feature selection from different data sets found in Mining Industry for identifying Tuberculosis (TB) disease, the novelty of proposal has not been validated.

It has been noted that the Radio-Frequency-Identification (RFID) technologies received significant attention in the healthcare industry by performing tasks such as tracking of medical equipment, hospital supplies, medication and patient information in an attractive proposition. Though the technological impacts are significant, it equally creates privacy concerns raised by RFID tag bearers. Rahman et al.(2017) attempted RFID based healthcare framework for addressing the privacy issues. The proposed framework consists of two components namely (i) Prisens – used as privacy preserving authentication protocol for sensing RFID tags for different identification and monitoring purposes. It includes drug usage monitoring, surgical instrument tracking, hospital personnel tracking and blood tracking (ii) HSAC – provides a privacy preserving healthcare service accessing mechanism for maintaining user's privacy while accessing various healthcare services. In addition, it follows the role based access mechanism for restricting the unauthorized access of private data that includes both structured and unstructured clinical information. The framework holds good for cheaper tags (tags with less computational capabilities) as well. Though the motivation behind their framework is to enhance the privacy of users in RFID based healthcare system, inculcating the adherence of different privacy levels for different service requests will add promising potential to their research efforts.

A cloud based distributed health information system framework that focuses on privacy and security aspects of medical information has been put forth by Sarkar (2017). The suggested framework applies set of security constraints and access control mechanisms that guarantee integrity, confidentiality, and privacy for medical data. It consists of various components such as electronic health records, care delivery organization, data analytics, and end users. A three tier sensitive system has been adopted for categorizing patient information stored in the electronic health records. Information such as name of the disease and its status, mental status of patient, biometric identifiers, and mediclaim numbers are ensured with Tier-0 sensitivity. The next level of security (Tier-1) has been applied to authenticate information such as date of birth, name, doctor's name etc. Information such as zip code, blood group, surgery name is authenticated in terms of Tier-2 sensitivity. To prevent privacy and security of patient data from malicious hackers, the framework adopts a two level security mechanism (Level-1 and Level-2). Accordingly, level-1 security relates to the authorization of user towards accessing patient data in the hospital by providing both temporary user-id and patient-id. To access the patient information at the inter-hospital and intra-hospital levels, an OTP based level-2 security mechanism has been enforced. Though the framework applies a set of security constraints for maintaining the privacy and confidentiality of medical data, the work mainly relies upon conceptual level only. Hence it requires successful implementation with the aid of necessary infrastructure to claim the validity.

A big data analytical framework for Voice Pathology Assessment (VPA) has been proposed by Hossain and Muhammad (2016). The framework facilitates the process of extracting, storing, processing and classifying the speech signals from patients. A cloud server receives speech signals from various hospitals and processes them by using two feature extraction techniques namely, MPEG-7 and Interlaced Derivative Pattern (IDP). Machine learning algorithms such as support vector machine, extreme learning machine, and Gaussian mixture model are applied to classify the signal as normal or pathological. Though the novel framework is a pioneering attempt to process voice signals of patient for building healthcare monitoring system, the proposed work does not consider the integration of other sources of big data into voice signals for performing the pathological assessment.

Pramanik et al. (2017) carried out critical analysis on the recent advancements in healthcare systems with significant focus on usage of smart system technologies. They advocated a conceptual framework for a big data enabled smart healthcare system towards providing ubiquitous healthcare solutions at reduced cost with increased intelligence. It includes layers such as i) Data source layer for handling structured, unstructured and semi-structured sources of data ii) Data analytics layer for performing big data intensive computation, management and visualization iii) Smart service layer for facilitating services such as data monitoring, privacy and security agreement between consumers and service providers. Besides the layer offers smart service infrastructure with the aid of devices, software and domain entities iv) Knowledge discovery layer added extensive functionalities such as prediction of entity needs, planning and estimation, evaluation and modeling of healthcare service mechanisms. Overall the framework creates opportunities for healthcare organizations in delivering intelligent smart system services. Their research offers a conceptual framework for a big data enabled smart healthcare system and leads to

interdisciplinary research directions. Though the framework is the application of three technical branches namely Intelligent Agents, Machine Learning and Text Mining in the context of smart health, the outcome in terms of implementation has not been evaluated in their work.

The usage of Internet of Things (IoT) in smart health demands healthcare services to enforce strict security and data quality in order to preserve the confidentiality and sensitivity of medical data. To support this initiative in large scale heterogeneous smart health environments, Sicari et al.(2017) proposed a flexible policy enforcement framework. The framework uses policy enforcement point (PEP), policy decision point (PDP) and policy administration point (PAP) to cater the needs of various stakeholders (Doctor, Patient, and Visitor). PEP is responsible for intercepting the requests posted by the users for accessing the resources and forwards it to PDP. The PDP on receiving the requests from PEP evaluates it against the authorization policies to arrive at a decision (approved or rejected) and notify it to PEP for granting or denial of access. The entire administration for authorizing policies including runtime policy updates has been carried out by PAP. Smart nodes that include wearable sensors, RFID's and monitors are also performing request initiations on behalf of the aforementioned stakeholders. Though the proposed framework offers a prototype based centralized policy enforcement mechanism for ensuring confidentiality of medical data, it needs further extension to support distributed security policy management solution for coordinating various smart health environments.

A big data framework for the provision of personalized real-time health care services on the basis of context-aware monitoring technology has been put forth by Forkan et al. (2015). The framework facilitates the analysis of big data inside a cloud environment to find patient specific anomalies from large amount of data. With the advent of cloud, continuously generated big data from heterogeneous context of various living system have been gathered and a 2-step learning methodology has been advocated. In the first step, a Map-Reduce based apriori algorithm has been applied to find out the correlation among context attributes and the threshold values of vital parameters of patient data. As a result, patient specific association rules are mined. In the second step, supervised learning algorithms such as Multi-Layer perception network, Decision Tree, Bayes Network are implemented over the generated association rules for performing context-aware decision. The framework has been implemented in various operational environments such as Single mode (execution of learning algorithm in single core), Pseudo-distributed mode (execution of learning algorithm in parallel), and Cloud mode (execution of learning algorithm in Amazon Elastic MapReduce). In a nut shell, the proposed framework leverages the advantages of context-aware computing, remote-monitoring, cloud computing, machine learning and big data. As a cloud based big data framework that utilizes machine learning approaches to detect anomaly situation of patient, it efficiently performs context-aware decision making in offering personalized healthcare services. However, the process of generating, pre-processing and transmission of sensor data from the patient context has not been perceived.

Recently, the impact of big data analytics in developing novel solution for protein structure prediction problem received significant attention among research community. Proteins are the key functional units in living organism and serve as hormones, receptors, storage, enzymes and as transporters of particles in human body and responsible for biochemical reactions. The huge volume, unstructured format, and ever-changing nature of protein sequence residues found in the Protein Data Bank (PDB) prompt the need of big data intensive computational framework for protein structure prediction problem.

Dencelin and Ramkumar (2016) proposed a big data framework using Apache Spark project that adopts Resilient Distributed Data sets (RDD) for secondary structure of proteins. The proposed framework performs machine learning based Multilayer Perceptron Algorithm using different set of input features and network parameters in distributed computing environment. The input PDB data sets (protein sequences) are classified into three secondary structures namely α-helix, β-sheet and coils. They have analysed the classifier accuracy with varied set of parameters by changing the input dataset features and perceptron parameters. In another work (Xavier and Thirunavukarasu, 2017), they have advocated an Ensemble based Random Forest algorithm to classify the protein structure in a Spark based distributed computing environment. Two bench marking datasets, RS126 and CB513 were taken into account for the implementation and improved accuracy has been obtained. Both the frameworks provide a novel opportunity to obtain a scalable, fault tolerance, efficient and reliable computing performance on Linux clusters. When these frameworks consider deep learning based algorithm for protein structure prediction, it will enhance the accuracy outcome of their proposed work.

The extensive research efforts discussed so far in developing various healthcare frameworks shows the importance of adopting big data based platforms and analytical techniques for reaping quality knowledge and disseminating it to the diversified healthcare stakeholders. In Table 2, we summarize the discussed healthcare frameworks under different features such as data source utilized, analytical capability and application areas.

The next section elaborates the implications of various big data tools that are useful in attaining significant artefacts in the healthcare lifecycle based on big data.

4. Implications of big data tools in developing healthcare frameworks

Though diversified big data frameworks are designed towards meeting specific healthcare objectives, they themselves orient well for adopting standard architectural guidelines for performing activities such as data gathering, pre-processing, data analysis, interpretation, and visualization. Due to domain-specific nature of big data healthcare framework, professionals such as data scientist should take utmost care in selecting appropriate tools (Philip Chen and Zhang, 2014; Sukumar et al., 2015) to be used at each level of the framework design and implementation. In this section, we insight the usage of various big data tools that play significant role in executing tasks such as integration of data, injecting intelligence, searching and indexing, stream data processing, and data visualization.

4.1. Data integration tools

The continuous growth in the volume and velocity of healthcare data with diversified data types demands the necessity of utilizing the services of data integration tools for aggregating data from disparate sources.

Pentaho (2017) is a big data analytical platform that provides end-to-end data integration to support users for analyzing data from disparate sources such as relational databases, Hadoop distributions, NoSQL stores and enterprise applications. It also provides a flexible user interface for creating visual data flows to perform transformation and integration of data.

Palantir (2017) is a data integration tool that rapidly fuses data from disparate sources such as medical device outputs and medical codes. Further, it enables analytical techniques to develop models for tracking sequence of procedures and clinical data metrics to manage healthcare diagnosis. Ayata (2017) efficiently brings

Table 2
Summary of healthcare big data frameworks.

Authors	Framework Name	Data Source	Analytical Capability	Application Area	Highlights
Raghupathi and Raghupathi (2014)	Generic Conceptual framework.	Geographically disparate data sources with multiple formats.	Queries and Reports.	General Healthcare architecture	Advocates a benchmarking framework for analyzing Healthcare Big Data
Chawla and Davis (2013)	A patient centric personalized healthcare framework	Electronic medical records, patient experiences, and histories	Collaborative filtering	Personalized healthcare	Data driven computational aid for personalized healthcare
Kim et al. (2014)	Big Data framework for Ubiquitous healthcare system	Continuous time series data such as ECG, respiration, and SpO2	Map-Reduce programming model	Personalized healthcare	Big data platform for processing vital signs
Fang et al. (2016)	Health informatics processing pipeline framework	Electronic Health Records, Public health, Genomic, Behavioural data	Feature selection and machine learning algorithms	Decision support system for practitioners	Systematic data processing pipeline for generic big health informatics
Youssef (2014)	Framework for secure health information system	Electronic Health Records	Map-Reduce	Security and privacy of healthcare data	Integrating mobile, cloud and big data technologies to provide personalized healthcare
Lin et al. (2015)	A cloud based framework for home diagnosis service	Patient data, patient profile and clinical data	Formal Concept Analysis (FCA)	Self-care diagnosis	Symptom-based medical record retrieval according to the user query
Legaz-Garca et al (2016)	A semantic web technology framework for managing and reusing clinical archetypes	Electronic Health Records	Ontology building through OWL.	Patient classification system based on clinical criteria	Integration of semantic resources with Electronic Health Records
Sakr and Elgammal (2016)	Smart Health Framework	Patient data from sources such as Hospital Information System, Radiology and laboratory information systems	Predictive modeling and pattern matching techniques	Data analytics for smart healthcare applications	Enhancing healthcare services by integrating sensor technologies, cloud computing and big data analytics
Mahmud et al. (2016)	Data analytics and visualization framework for health-shocks prediction	Healthcare data set focusing on context such as socioeconomic, cultural and geographical conditions	Predictive modeling using Fuzzy rule summarization	Public health services	Integrating cloud computing services with geographical information system
Jokonya (2014)	Big data integrated framework for forecasting and controlling diseases in Mining Industry	Epidemics data about infectious diseases	Predictive analytics	Disease forecasting and prevention	Holistic and structured approach to understanding the complex linkage of HIV/AIDS, tuberculosis and silicosis in mining industry
Rahman et al. (2017))	Framework for preserving privacy in RFID based healthcare systems	Data generated from RFID tags	Privacy preserving techniques	Secure healthcare services	Improved privacy in RFID based healthcare system
(Sarkar, 2017	Conceptual framework for secured distributed health information system	Electronic Health Records	Provisioning security constraints and access control mechanisms	Secure healthcare system	Distributed framework for two-level security mechanism
Hossain and Muhammad (2016)	Voice Pathology Assessment Framework	Voice Signals	Classification Techniques	To classify the signal as normal or pathological.	A cloud based big data analytical framework for handling unstructured data
Pramanik et al. (2017)	Big Data enabled smart healthcare system framework	EHR, diagnosis report, social media data, surveillance data, biometric data	Provisioning smart healthcare services through service oriented infrastructure	Smart system technologies for state of the art healthcare system	Coupling of healthcare knowledge discovery mechanism with smart service infrastructure
Sicari et al. (2017)	IoT based policy enforcement framework for smart health	Biological parameters of patient, environmental data, RFID and instruments generated data	Offering policy based access control mechanism for availing healthcare resources	Smart health applications to prevent security threats in large scale heterogeneous scenarios	Enforcing policy primitives for smarter healthcare infrastructures
Forkan et al. (2015)	Cloud based big data framework for personalized patient care through context aware monitoring	Patient profile data, medical records, activity logs, vital signs and context cum environmental data generated from sensors	A two-step learning methodology that includes correlation analysis of context attributes using association rule mining followed by supervised learning strategy	Mining trends and patterns in patient data to support individual centric healthcare services	Personalized healthcare services through context aware decision making approach
Dencelin and Ramkumar (2016)	Big Data Framework for Structure Prediction of Proteins using Multi-Layer Perceptron	PDB dataset	Neural network based classification technique	Drug Design	Advocates a Spark based Big Data Framework for Secondary structure prediction of protein.
Xavier and Thirunavukarasu (2017)	Big Data Framework for Structure Prediction of Proteins using Ensemble Learning	PDB dataset	Distributed Tree Based Ensemble learning Technique	Drug Design	Advocates a Distributed Spark based Framework with improved accuracy

together structured and unstructured (videos, images, text, sound) healthcare data, mathematical models, business rules to build predictive and prescriptive models.

As data integration software, Attunity (2017) ingests data from disparate data sources in an efficient, cost effective and scalable fashion. It integrates data from major sources such as data warehouse, Hadoop and cloud platforms in a rapid manner without manual coding. Informatica (2017) offers a wide range of data management services that includes analytics, integration and governance of healthcare data for improved patient care. It accesses and transforms both clinical and administrative data that conforms to HIPAA (Health Insurance Portability and Accountability Act) and HL7 standards into usable format. Jitterbit (2017) offers a single-secure platform for healthcare organizations to access clinical and workflow data by merging unstructured data with structured data in multiple standard formats. The data received from sources such as proprietary EHR systems can be transformed for comprehensive analysis.

4.2. Scalable searching and processing tools

Since large volume of clinical notes and unstructured text are commonly used by the physicians in the healthcare domain, there is an immense need for searching and indexing tools for performing optimized full-text search capability of clinical data. These tools are utilized for effective distributed text management and indexing large volume of data in file system such as HDFS (Hadoop Distributed File System).

Apache Lucene (2017) is a scalable, high-performance indexing system that offers powerful and accurate full-text search facility for variety of applications across different platforms. Google Dremel (Melnik et al., 2010) is a distributed system for interactively querying large data sets and supports nested data with column storage representation. It uses multi-level execution trees for query processing. Apache Drill is the Open Source implementation of Google Dremel.

Cloudera Impala (2017) offers high-performance, low-latency SQL queries on data stored in Apache Hadoop file formats. Impala integrates with the Apache Hive meta-store database to share databases and tables between them. It pioneers the use of Parquet file format, a columnar storage layout optimized for large-scale queries. Dryad is a general purpose distributed execution engine (Isard et al., 2007) that was designed to support a wide variety of parallel applications such as relational queries, large scale matrix computations and text-processing tasks. It supports scalability by extending its processing capabilities from very small to large clusters. Application represented in the form of a data flow graph gets executed on a set of available computers, communicating through files, TCP pipes, and shared-memory FIFOs.

4.3. Machine learning tools

The healthcare industry is keen in availing the applications of machine learning tools to transform the abundant medical data into actionable knowledge by performing predictive and prescriptive analytics in view of supporting intelligent clinical activities.

Apache Mahout (2017) is an open source machine learning library that sits on top of Hadoop to facilitate the execution of scalable machine learning algorithms for a distributed environment. It offers techniques such as Recommendation, Classification, and Clustering. Mahout applications include pattern mining, user interest modeling, and personalization. Skytree (2017) is a general purpose machine learning platform that uses artificial intelligence to produce sophisticated algorithms for performing advanced analytics. It has the ability to process massive datasets (structured and unstructured) in an accurate manner without down sampling. Few of its use cases are recommendation systems, anomaly/outlier identification, predictive analytics, clustering and market segmentation, and similarity search.

Karmasphere (2017) creates a big data platform that mines and analyzes the web, mobile, sensor and social media in Hadoop. It provides a graphical environment that supports navigation through big data of any variety and spot trends and patterns in it. Karmasphere had been acquired by FICO in 2014. BigML, 2017 is a scalable and programmable machine learning platform that provides several tools to perform machine learning tasks such as classification, regression, cluster analysis, anomaly detection and association discovery. It seamlessly integrates the features of machine learning with cloud infrastructure to build cost-effective applications with high scalability, flexibility, and reliability.

4.4. Real-time and stream data processing tools

Advances in IoT and sensor devices found in healthcare domain prompts the data processing from diversified data sources to be carried out in a real-time manner. The on-the-fly analysis of healthcare data enables the system to make better decisions for personalizing patient oriented services.

Apache Storm (2017) is a real-time data platform for processing limitless streaming data that boasts capability to integrate seamlessly with existing queuing and database technologies to process over a million tuples per second per node. Its applications include real-time analytics, interactive operation system, online machine learning, and ETL. S4 (Simple Scalable Streaming System) is a distributed stream processing engine (Neumeyer et al., 2010) that allows programmers to develop applications for processing continuous unbounded streams of data. Some of the key properties such as robustness, decentralization, scalability, cluster management, and extensibility have been offered by S4.

SQLstream Blaze (2017) is a streaming analytics platform that acquires data from all available sources in all formats and at all speeds (Unstructured, structured, local, distributed, in-motion, at-rest, live or historical). The Blaze has the potential to scale up its data handling capability to millions of records per second per CPU core with 1–5 ms latency. Splunk (2017) is a real-time and intelligent big data platform that enables organizations to gain operational intelligence from machine data for real-time insights. It performs indexing of structured /unstructured machine-generated data, real-time searching and reporting analytical results.

Apache Kafka (2017) is a distributed streaming platform used for building real-time streaming data pipelines (reliable transfer of data between systems or applications) and applications (transform or act on streams of data). It uses four core APIs (Producer, Consumer, Streams, and Connector) to facilitate services such as message passing, storage and stream processing. SAP Hana (2017) an in-memory analytics platform that provides real-time analysis along with support for various capabilities such as database services, advanced analytics, application development, data access, administration, and openness.

4.5. Visual data analytical tools

Data visualization tools in healthcare helps to identify patterns, trends and deviations that include outliers, clusters, association discovery and time series analysis for improving clinical healthcare delivery and public health policy.

Jaspersoft (2017) is a scalable big data analytical platform that supports effective decision making with the help of interactive

Table 3
Big Data Tools for Healthcare Ecosystem.

Task	Tools	Merits & its applications
Data Integration	Pentaho (2017)	Tool for performing knowledge discovery process from a scalable environment in a robust and flexible manner.
	Palantir (2017)	Assist decision makers for highlighting the process insights by uncovering treatment options and improving the standard of patient care.
	Ayata (2017)	Performs exclusive prescriptive analytics from large amount of data towards helping organizations for making smarter decisions.
	Attunity (2017)	Modern data integration platform for performing automated data pipelines at a faster rate from different architectures such as cloud and data lakes.
	Informatica (2017)	Provides enterprise information management solutions over data from diversified sources
	Jitterbit (2017)	Facilitates integration of data from Saas based cloud services and on-premise applications in an intelligent manner.
Searching and processing	Apache Lucene (2017)	High performance, full-featured text search engine for performing full-text search across different platforms.
	Google Dremel Melnik et al. (2010)	Complements Map/Reduce based computations supported by Hadoop for processing nested data with high scalability.
	Cloudera Impala (2017)	Executes low latency and high concurrency analytical queries on top of Hadoop.
	Dryad Isard et al. (2007)	Provides high performance distributed execution engine with good programming constructs.
Machine Learning	Apache Mahout (2017)	Offers distributed machine learning library for processing scalable mining algorithms.
	Skytree (2017)	Tool for performing machine learning and advanced analytics of massive data sets at high speed.
	Karmasphere (2017)	Big data workspace tool for discovering pattern insights from large volume of data stored on hadoop clusters.
	BigML (2017)	Platform for offering solutions to big data use-cases through predictive analytics.
Stream Data Processing	Apache Storm (2017)	Scalable and flexible real-time computation system for processing massive amount of data.
	S4 Neumeyer et al. (2010)	Tool for processing unbounded stream data in a distributed, scalable and fault-tolerant manner.
	SQLstream Blaze (2017)	Supports creation of distributed streaming applications that deliver data ingestion, integration, and analytics in real time.
	Splunk (2017)	Highly scalable tool to collect and harness machine data. Ability to scale from laptop to datacentre.
	Apache Kafka (2017)	A distributed, high throughput stream data processing tool for facilitating publish-subscribe messaging system.
	SAP Hana (2017)	Tool for in-memory computing of stream data for real-time analytics.
Visual Data Analytics	Jaspersoft (2017)	Provides interactive visual analytics at large scale by extracting information from one or more data sources.
	Tableau (2017)	Provides faster, highly interactive dashboards to project extracted patterns.
	Qlik (2017)	Explores clinical and operational data through visual analytics for discovering insights.

reports, analytics, and dashboards. It provides fast data visualization on storage platforms such as MongoDB, Cassandra, Redis, Riak, and CouchDB. Tableau (2017) has the ability to transform large, complex data sets into intuitive pictures. It combines advances in database and computer graphics technology to analyze huge datasets with limited computing resources. Qlik (2017) enables healthcare organizations to explore clinical, financial and operational data through visual analytics to discover insights that lead to improvements in care, reduced costs and delivering higher value to patients.

Table 3 accounts various big data tools with their merits that facilitate the execution of specified tasks in the healthcare ecosystem.

The above-discussed tools are helpful to the researchers for deploying effective healthcare frameworks that facilitate end-to-end healthcare solutions by improving patient outcomes with the advent of big data.

5. Conclusion

Framework based solutions always cater to the comprehensive requirement of various stakeholders involved in the healthcare domain. With the impact of big data, healthcare domain was revamped and offer intensive solutions for handling diversified big data sources that range from patient health records to medical images. This paper reviews various research attempts in establishing healthcare frameworks and summarizes their significant outcomes. The summary of contributions by various researchers highlights the data source utilized, adopted analytical techniques and other features. At the end, the implication of various big data tools in developing healthcare framework is also extensively studied.

References

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K., 2015. Efficient machine learning for big data: a review. Big Data Res. 2, 87–93. https://doi.org/10.1016/j.bdr.2015.04.001.

Andreu-Perez, J., Poon, C.C.Y., Merrifield, R.D., Wong, S.T.C., Yang, G.Z., 2015. Big Data for Health. IEEE J. Biomed. Heal. Informat. 19, 1193–1208. https://doi.org/10.1109/JBHI.2015.2450362.

Bouhriz, M., Chaoui, H., 2015. Big data privacy in healthcare moroccan context. Procedia Comput. Sci. 63, 575–580. https://doi.org/10.1016/j.procs.2015.08.387.

Chawla, N.V., Davis, D.A., 2013. Bringing big data to personalized healthcare: a patient-centered framework. J. Gen. Intern. Med. 28, 660–665. https://doi.org/10.1007/s11606-013-2455-8.

Dencelin, L.X., Ramkumar, T., 2016. Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures. Biomed. Res. 2016, S166–S173.

Dimitrov, D.V., 2016. Medical internet of things and big data in healthcare. Healthc. Inform. Res. 22, 156–163. https://doi.org/10.4258/hir.2016.22.3.156.

Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C., Iyengar, S.S., 2016. Computational Health Informatics in the Big Data Age: a survey. ACM Comput. Surv. 49, 1–36. https://doi.org/10.1145/2932707.

Forkan, A.R.M., Khalil, I., Ibaida, A., Member, Z.T., 2015. BDCaM: Big Data for Context-Aware monitoring-a personalized knowledge discovery framework for assisted healthcare. IEEE Trans. Cloud Comput. https://doi.org/10.1109/TCC.2015.2440269.

Hossain, M.S., Muhammad, G., 2016. Healthcare big data voice pathology assessment framework. IEEE Access 4, 7806–7815. https://doi.org/10.1109/ACCESS.2016.2626316.

Isard, M., Budiu, M., Yu, Y., Birrell, A., Fetterly, D., 2007. Dryad: distributed data-parallel programs from sequential building blocks. ACM SIGOPS Oper. Syst. Rev. 59–72. https://doi.org/10.1145/1272998.1273005.

Jee, K., Kim, G.H., 2013. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc. Inform. Res. 19, 79–85. https://doi.org/10.4258/hir.2013.19.2.79.

Jokonya, O., 2014. Towards a big data framework for the prevention and control of HIV/AIDS, TB and silicosis in the mining industry. Procedia Technol. 16, 1533–1541. https://doi.org/10.1016/j.protcy.2014.10.175.

Kim, T.W., Park, K.H., Yi, S.H., Kim, H.C., 2014. A big data framework for u-healthcare systems utilizing vital signs. In: Proc. – 2014 Int. Symp. Comput. Consum. Control. IS3C 2014 494–497, doi:10.1109/IS3C.2014.135.

Kupwade Patil, H., Seshadri, R., 2014. Big data security and privacy issues in healthcare. IEEE Int. Congr. Big Data pp. 762–765. doi:10.1109/BigData.Congress.2014.112.

Legaz-Garca, M.D.C., Martinez-Costa, C., Menarguez-Tortosa, M., Fernandez-Breis, J. T., 2016. A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. Knowledge-Based Syst. 105, 175–189. https://doi.org/10.1016/j.knosys.2016.05.016.

Liang, Y., Kelemen, A., 2016. Big data science and its applications in health and medical research: challenges and opportunities. J. Biom. Biostat. 7, 1–9. https://doi.org/10.4172/2155-6180.1000307.

Lin, W., Dou, W., Zhou, Z., Liu, C., 2015. A cloud-based framework for Home-diagnosis service over big medical data. J. Syst. Softw. 102, 192–206. https://doi.org/10.1016/j.jss.2014.05.068.

Mahmud, S., Iqbal, R., Doctor, F., 2016. Cloud enabled data analytics and visualization framework for health-shocks prediction. Futur. Gener. Comput. Syst. 65, 169–181. https://doi.org/10.1016/j.future.2015.10.014.

Mancini, M., 2014. Exploiting big data for improving healthcare services. J. E-Learning Knowl. Soc. 10, 23–33.

Martin-Sanchez, F., Verspoor, K., 2014. Big data in medicine is driving big changes. Yearb. Med. Inform. 9, 14–20. https://doi.org/10.15265/IY-2014-0020.

Melnik, S., Gubarev, A., Long, J.J., Romer, G., Shivakumar, S., Tolton, M., Vassilakis, T., 2010. Dremel: Interactive Analysis of Web-Scale Datasets. In: 36th Int. Conf. Very Large Data Bases pp. 330–339. doi:10.1145/1953122.1953148.

Merelli, I., Pérez-Sánchez, H., Gesing, S., D'Agostino, D., 2014. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res. Int. 2014. https://doi.org/10.1155/2014/134023.

Neumeyer, L., Robbins, B., Nair, A., Kesari, A., 2010. S4: Distributed stream computing platform. In: Proc. – IEEE Int. Conf. Data Mining, ICDM pp. 170–177. doi:10.1109/ICDMW.2010.172.

Philip Chen, C.L., Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf. Sci. (Ny) 275, 314–347. https://doi.org/10.1016/j.ins.2014.01.015.

Pramanik, M.I., Lau, R.Y.K., Demirkan, H., Azad, M.A.K., 2017. Smart health: big data enabled health paradigm within smart cities. Expert Syst. Appl. 87, 370–383. https://doi.org/10.1016/j.eswa.2017.06.027.

Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. Heal. Inf. Sci. Syst. 2, 3. https://doi.org/10.1186/2047-2501-2-3.

Rahman, F., Bhuiyan, M.Z.A., Ahamed, S.I., 2017. A privacy preserving framework for RFID based healthcare systems. Futur. Gener. Comput. Syst. 72, 339–352. https://doi.org/10.1016/j.future.2016.06.001.

Sakr, S., Elgammal, A., 2016. Towards a comprehensive data analytics framework for smart healthcare services. Big Data Res. 4, 44–58. https://doi.org/10.1016/j.bdr.2016.05.002.

Sarkar, B.K., 2017. Big data for secure healthcare system : a conceptual design. Complex Intell. Syst. 3, 133–151. https://doi.org/10.1007/s40747-017-0040-1.

Sedig, K., Ola, O., 2014. The challenge of big data in public helth: an opportunity for visual analytics. Online J. Public Health Inform. 5, 1–21. https://doi.org/10.5210/ojphi.v5i3.4933.

Sicari, S., Rizzardi, A., Grieco, L.A., Piro, G., Coen-Porisini, A., 2017. A policy enforcement framework for Internet of Things applications in the smart health. Smart Heal. 3–4, 39–74. https://doi.org/10.1016/j.smhl.2017.06.001.

Sukumar, S.R., Natarajan, R., Ferrell, R.K., 2015. Quality of big data in health care. Int. J. Health Care Qual. Assur. 28, 621–634. https://doi.org/10.1108/IJHCQA-07-2014-0080.

Wang, L., Alexander, C.A., 2015. Big data in medical applications and health care. Am. Med. J. 6, 1–8. https://doi.org/10.3844/amjsp.2015.1.8.

Wu, P.-Y., Cheng, C.-W., Kaddi, C.D., Venugopalan, J., Hoffman, R., Wang, M.D., 2017. Omic and electronic health record big data analytics for precision\Nmedicine. Ieee Trans. Biomed. Eng. 64, 263–273. https://doi.org/10.1109/TBME.2016.2573285.

Xavier, L., Thirunavukarasu, R., 2017. A distributed tree-based ensemble learning approach for efficient structure prediction of protein. Int. J. Intell. Eng. Syst. 10, 226–234. https://doi.org/10.22266/ijies2017.0630.25.

Youssef, A.E., 2014. A Framework for secure healthcare systems based on big data analytics in mobile cloud computing environments. Int. J. Ambient Syst. Appl. 2, 1–11. https://doi.org/10.5121/ijasa.2014.2201.

Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., 2017. Machine learning on big data: opportunities and challenges. Neurocomputing 237, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026.

Web References

Pentaho, <http://www.pentaho.com/solutions/healthcare/> (accessed 01.09.17.).

Palantir, <https://www.palantir.com/solutions/healthcare-delivery/> (accessed 01.09.17.).

Ayata, <http://ayata.com/prescriptive-analytics/> (accessed 01.09.17.).

Attunity, <https://www.attunity.com/solutions/hadoop-big-data/data-ingestion-hadoop/> (accessed 20.10.17.).

Informatica, <https://www.informatica.com/in/solutions/industry-solutions/healthcare.html/> (accessed 20.10.17.).

Jitterbit, <https://www.jitterbit.com/solutions/integration-solutions-by-industry/healthcare-life-sciences/> (accessed 20.10.17).

Apache Lucene, <https://lucene.apache.org/core/> (accessed 09.09.2017.).

Cloudera Impala, <https://www.cloudera.com/documentation/enterprise/5-3-x/topics/impala_intro.html> (accessed 10.08.2017).

Apache Mahout <http://mahout.apache.org/> (accessed 20.08.017.).

Skytree, <http://www.skytree.net/products/> (accessed 20.08.2017.).

Karmasphere (FICO), <http://www.fico.com/en/communications/patient-adherence#overview> (accessed 09.09.2017.).

BigML, <https://bigml.com/> (accessed 20.08.2017.).

Apache Storm, <http://storm.apache.org/index.html> (accessed 10.08.17.).

SQLstream Blaze, <http://sqlstream.com/capabilities/> (accessed 15.08.17.).

Splunk, <https://www.splunk.com/en_us/solutions/industries/healthcare.html> (accessed 01.09.17.).

Apache Kafka, <https://kafka.apache.org/intro> (accessed 1.08.17.).

SAP Hana <https://www.sap.com/india/industries/healthcare.html> (accessed 1.08.17.).

Jaspersoft, <https://www.jaspersoft.com/features> (accessed 15.08.17.).

Tableau, <https://www.tableau.com/solutions/topic/healthcare> (accessed 01.09.17.).

Qlik, <http://www.qlik.com/en-in/solutions/industries/healthcare> (accessed 09.09.17.).