ijergs
Open Access
International Journal of Engineering Research and Generic Science (IJERGS)
Available Online at www.ijergs.in
ISSN: 2455 - 1597
Volume -2, Issue-5, September - October 2016, Page No. 46 - 55

A Review: Big Data and Analytics in Health Care

L.Jude Ashwin Jeyakumar[1], Dr. K.Mahalakshmi[2], Mrs. R. Keerthika[3], Mrs. A.Nagajothi[4]
[1]Student, Department of Information Technology
[2]Professor, Department of Information Technology
[3]Asst Professor, Department of Information Technology
[4] Asst Prof, Department of Computer Science and Engineering
[1, 2, 3, 4]Karpagam College of Engineering, Coimbatore Tamil Nadu, India
E-Mail Id: prof.dr.mlk@gmail.com

Abstract
Big data analytics is a growth area with the potential to provide useful insight in healthcare. Whilst many dimensions of big data still present issues in its use and adoption, such as managing the volume, variety, velocity, veracity, and value, the accuracy, integrity, and semantic interpretation are of greater concern in clinical application. However, such challenges have not deterred the use and exploration of big data as an evidence source in healthcare. This drives the need to investigate healthcare information to control and reduce the burgeoning cost of healthcare, as well as to seek evidence to improve patient outcomes. Objective is to describe the promise and potential of big data analytics in healthcare. This paper describes the nascent field of big data analytics in healthcare, discusses the benefits, outlines an architectural framework and methodology, describes examples reported in the literature, briefly discusses the challenges, and offers conclusions. This paper provides a broad overview of big data analytics for healthcare. Finally the paper is concluded that big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

Keywords: Big data, Analytics, Hadoop, Healthcare, Framework, Methodology.

1. Introduction

The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance & regulatory requirements, and patient care [1]. While most data is stored in hard copy form, the current trend is toward rapid digitization of these large amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data (known as 'big data') hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management [2-5]. Reports say data from the healthcare system alone reached, in 2011, 150 exabytes. At this rate of growth, big data for healthcare will soon reach the zettabyte (1021 gigabytes) scale and, not long after, the yottabyte (1024 gigabytes) [6]. By definition, big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/ or hardware; nor can they be easily managed with traditional or common data management tools and methods [7]. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed [7]. The totality of data related to patient healthcare and wellbeing make up "big data" in the healthcare industry. For the big data scientist, there is, amongst this vast amount and array of data, opportunity. By discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve care, save lives and lower costs. Thus, big data analytics applications in healthcare take advantage of

the explosion in data to extract insights for making better informed decisions [10-12], and as a research category are referred to as, no surprise here, big data analytics in healthcare [13-15]. When big data is synthesized and analyzed— and those aforementioned associations, patterns and trends revealed—healthcare providers and other stakeholders in the healthcare delivery system can develop more thorough and insightful diagnoses and treatments, resulting, one would expect, in higher quality care at lower costs and in better outcomes overall [12]. The potential for big data analytics in healthcare to lead to better outcomes exists across many scenarios, for example: by analyzing patient characteristics and the cost and outcomes of care to identify the most clinically and cost effective treatments and offer analysis and tools, thereby influencing provider behavior; applying advanced analytics to patient profiles (e.g., segmentation and predictive modeling) to proactively identify individuals who would benefit from preventative care or lifestyle changes; broad scale disease profiling to identify predictive events and support prevention initiatives; collecting and publishing data on medical procedures, thus assisting patients in determining the care protocols or regimens that offer the best value; identifying, predicting and minimizing fraud by implementing advanced analytic systems for fraud detection and checking the accuracy and consistency of claims; and, implementing much nearer to real-time, claim authorization; creating new revenue streams by aggregating and synthesizing patient clinical records and claims data sets to provide data and services to third parties, for example, licensing data to assist pharmaceutical companies in identifying patients for inclusion in clinical trials. Many payers are developing and deploying mobile apps that help patients manage their care locate providers and improve their health. Via analytics, payers are able to monitor adherence to drug and treatment regimens and detect trends that lead to individual and population wellness benefits [12, 16-18]. This article provides an overview of big data analytics in healthcare as it is emerging as a discipline. First, we define and discuss the various advantages and characteristics of big data analytics in healthcare. Then we describe the architectural framework of big data analytics in healthcare. Third, the big data analytics application development methodology is described. Fourth, the challenges are identified. Lastly, we offer conclusions and future directions.

Big data analytics in healthcare

Health data volume is expected to grow dramatically in the years ahead [6]. In addition, healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in today's healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits [18]. Big data encompasses such characteristics as variety, velocity and, with respect specifically to healthcare, veracity. Existing analytical techniques can be applied to the vast amount of existing (but currently unanalyzed) patient-related health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient.

Advantages to healthcare

By digitizing, combining and effectively using big data, healthcare organizations ranging from single-physician offices and multi-provider groups to large hospital networks and accountable care organizations stand to realize significant

benefits [2]. Potential benefits include detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Numerous questions can be addressed with big data analytics. Big data could help to reduce waste and inefficiency in the following three areas:

Clinical operations: Comparative effectiveness research to determine more clinically relevant and cost-effective ways to diagnose and treat patients.

Research & development: 1) predictive modeling to lower attrition and produce a leaner, faster, more targeted R & D pipeline in drugs and devices.

Public health: Analyzing disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response.

Evidence-based medicine: Combine and analyze a variety of structured and unstructured data-EMRs, financial and operational data, clinical data, and genomic data to match treatments with outcomes, predict patients at risk for disease or readmission and provide more efficient care.

Genomic analytics: Execute gene sequencing more efficiently and cost effectively and make genomic analysis a part of the regular medical care decision process and the growing patient medical record.

Pre-adjudication fraud analysis: Rapidly analyze large numbers of claim requests to reduce fraud, waste and abuse.

Device/remote monitoring: Capture and analyze in real-time large volumes of fast-moving data from in-hospital and in-home devices, for safety monitoring and adverse event prediction.

Patient profile analytics: Apply advanced analytics to patient profiles (e.g., segmentation and predictive modeling) to identify individuals who would benefit from proactive care or lifestyle changes, for example, those patients at risk of developing a specific disease (e.g., diabetes) who would benefit from preventive care [14].

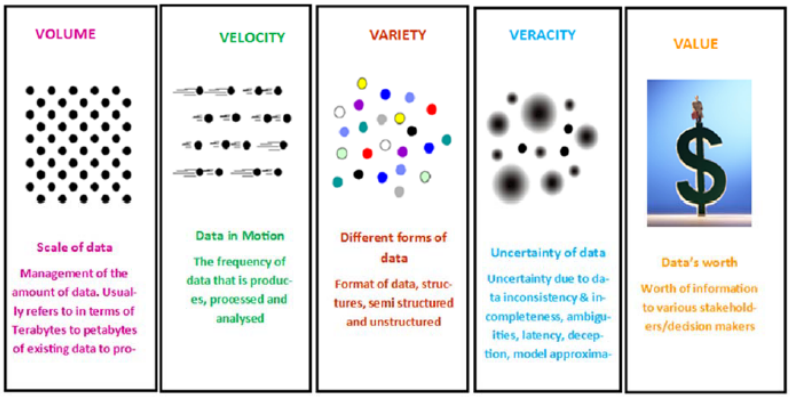

Figure 1: The 5 "Vs" of big data

These five dimensions (Fig. 1) play an important role in the issues facing big data analytics. Big data analytics involves the analysis of high volume data that stems from a variety of sources, including structured and unstructured data (Commonwealth of Australia, 2013). It also refers to the volume of the data sets within the data analysis and the velocity in which they are analysed. In recent years, the concept of big data has.

2. Architecture Framework

The conceptual framework for a big data analytics project in healthcare is similar to that of a traditional health informatics or analytics project. The key difference lies in how processing is executed. In a regular health analytics project, the analysis can be performed with a business intelligence tool installed on a stand-alone system, such as a desktop or laptop. Because big data is by definition large, processing is broken down and executed across multiple nodes. The concept of distributed processing has existed for decades. What is relatively new is its use in analyzing very large data sets as healthcare providers start to tap into their large data repositories to gain insight for making better-informed health-related decisions. Furthermore, open source platforms such as Hadoop/MapReduce, available on the cloud, have encouraged the application of big data analytics in healthcare. While the algorithms and models are similar, the user interfaces of traditional analytics tools and those used for big data are entirely different; traditional health analytics tools have become very user friendly and transparent.

Big data analytics tools, on the other hand, are extremely complex, programming intensive, and require the application of a variety of skills. They have emerged in an ad hoc fashion mostly as open-source development tools and platforms, and therefore they lack the support and user-friendliness that vendor-driven proprietary tools possess.

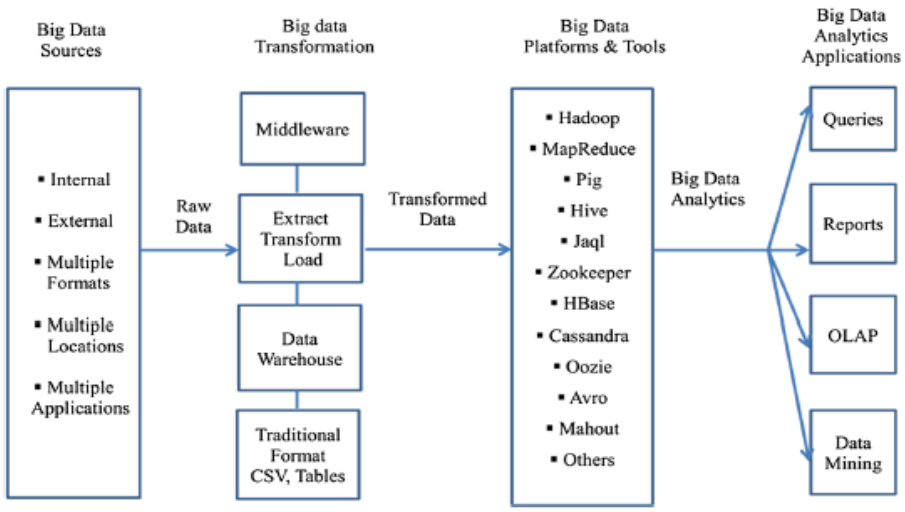

Figure 2: An applied conceptual architecture of big data analytics

As Figure 2 indicates, the complexity begins with the data itself. Big data in healthcare can come from internal (e.g., electronic health records, clinical decision support systems, etc.) and external sources (government sources, laboratories, pharmacies, insurance companies & HMOs, etc.), often in multiple formats (flat files, .csv, relationaltables, ASCII/text, etc.) and residing at multiple locations (geographic as well as in different healthcare providers' sites) in numerous legacy and other applications (transaction processing applications, databases, etc.). Sources and data types include:

1. Web and social media data: Clickstream and interaction data from Facebook, Twitter, LinkedIn,blogs, and the like. It can also include health plan websites, Smartphone apps, etc. [6].

2. Machine to machine data: readings from remote sensors, meters, and other vital sign devices [6].

3. Big transaction data: health care claims and other billing records increasingly available in semi-structured and unstructured formats [6].

4. Biometric data: finger prints, genetics, handwriting, retinal scans, x-ray and other medical images, blood pressure, pulse and pulse-oximetry readings, and other similar types of data [6].

5. Human-generated data: unstructured and semi-structured data such as EMRs, physicians notes, email, and paper documents [6].

For the purpose of big data analytics, this data has to be pooled. In the second component the data is in a 'raw' state and needs to be processed or transformed, at which point several options are available. A service oriented architectural approach combined with web services (middleware) is one possibility [17]. The data stays raw and services are used to call, retrieve and process the data. Another approach is data warehousing wherein data from various sources is aggregated and made ready for processing, although the data is not available in realtime. Via the steps of extract, transform, and load (ETL), data from diverse sources is cleansed and readied. Depending on whether the data is structured or unstructured, several data formats can be input to the big data analytics platform. In this next component in the conceptual framework, several decisions are made regarding the data input approach, distributed design, tool selection and analytics models. Finally, on the far right, the four typical applications of big data analytics in healthcare are shown. These include queries, reports, OLAP, and data mining.

Visualization is an overarching theme across the four applications. Drawing from such fields as statistics, computer science, applied mathematics and economics, a wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyze, and visualize big data in healthcare. The most significant platform for big data analytics is the open-source distributed data processing platform Hadoop (Apache platform), initially developed for such routine functions as aggregating web search indexes. It belongs to the class "NoSQL" technologies—others include CouchDB and MongoDB—that evolved to aggregate data in unique ways. Hadoop can serve the twin roles of data organizer and analytics tool. It offers a great deal of potential in enabling enterprises to harness the data that has been, until now, difficult to manage and analyze. Specifically, Hadoopmakes it possible to process extremely large volumes of data with various structures or no structure at all. But Hadoop can be challenging to install, configure and administer, and individuals with Hadoop skills are not easily found. Furthermore, for these reasons, it appears organizations are not quite ready to embrace Hadoop completely. The surrounding ecosystem of additional platforms and tools supports the Hadoop distributed platform. These are summarized in Table 1. Numerous vendors—including AWS, Cloudera, Hortonworks, and MapR Technologies—distribute open source Hadoop platforms [19]. While the development costs may be lower since these tools are open source and free of charge, the downsides are the lack of technical support and minimal security. In the healthcare industry, these are, of course, significant drawbacks, and therefore the trade-offs must be addressed. Additionally, these platforms/tools require a great deal of programming, skills the typical end-user in healthcare may not possess. Furthermore, considering the only recent emergence of big data analytics in healthcare, governance issues including ownership, privacy, security, and standards have yet to be addressed. In the next section we offer an applied big data analytics in healthcare methodology to develop and implement a big data project for healthcare providers.

Table 1 Platforms & tools for big data analytics in healthcare

Platform/Tool	Description
The Hadoop Distributed File System (HDFS)	HDFS enables the underlying storage for the Hadoop cluster. It divides the data into smaller parts and distributes it across the various servers/nodes.
Map Reduce	Map Reduce provides the interface for the distribution of sub-tasks and the gathering of outputs. When tasks are executed, Map Reduce tracks the processing of each server/node.
PIG and PIG Latin (Pig and Pig Latin)	Pig programming language is configured to assimilate all types of data (structured/unstructured, etc.). It is comprised of two key modules: the language itself, called Pig Latin, and the runtime version in which the Pig Latin code is executed.
Hive	Hive is a runtime Hadoop support architecture that leverages Structure Query Language (SQL) with the Hadoop platform. It permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical SQL statements
Jaql	Jaql is a functional, declarative query language designed to process large data sets. To facilitate parallel processing, Jaql converts "'high-level' queries into 'low-level'queries" consisting of Map Reduce tasks
Zookeeper	Zookeeper allows a centralized infrastructure with various services, providing synchronization across a cluster of servers. Big data analytics applications utilize these services to coordinate parallel processing across big clusters.
HBase	HBase is a column-oriented database management system that sits on top of HDFS. It uses a non-SQL approach.
Cassandra	Cassandra is also a distributed database system. It is designated as a top-level project modeled to handle big data distributed across many utility servers. It also provides reliable service with no particular point of failure ttp://en.wikipedia.org/wiki/Apache_Cassandra) and it is a NoSQL system.
Oozie	Oozie, an open source project, streamlines the workflow and coordination among the tasks.
Lucene	The Lucene project is used widely for text analytics/searches and has been incorporated into several open source projects. Its scope includes full text indexing and library search for use within a Java application.
Avro	Avro facilitates data serialization services. Versioning and version control are additional useful features.
Mahout	Mahout is yet another Apache project whose goal is to generate free applications of distributed and scalable machine learning algorithms that support big data analytics on the Hadoop platform.

3. Methodology

While several different methodologies are being developed in this rapidly emerging discipline, here we outline one that is practical and hands-on. Table 2 shows the main stages of the methodology. In Step 1, the interdisciplinary big data analytics in healthcare team develops a 'concept statement'. This is a first cut at establishing the need for such a project. The concept statement is followed by a description of the project's significance. The healthcare organization will note that there are trade-offs in terms of alternative options, cost, scalability, etc. Once the concept statement is approved, the team can proceed to Step 2, the proposal development stage. Here, more details are filled in. Based on the concept statement, several questions are addressed: What problem is being addressed? Why is it important and interesting to the healthcare provider? What is the case for a 'big data' analytics approach? (Because the complexity and cost of big data analytics are significantly higher compared to traditional analytics approaches, it is important to justify their use). The project team also should provide background information on the problem domain as well as prior projects and research done in this domain. Next, in Step 3, the steps in the methodology are fleshed out and implemented. The concept statement is broken down into a series of propositions. (Note these are not rigorous as they would be in the case of statistical approaches. Rather, they are developed to help guide the big data analytics process). Simultaneously, the independent and dependent variables or indicators are identified. The data sources, as outlined in Figure 2, are also identified; the data is collected, described, and transformed in preparation for for analytics. A very important step at this point is platform/tool evaluation and selection. The next step is to apply the various big data analytics techniques to the data. This process differs from routine analytics only in that the techniques are scaled up to large data sets. Through a series of iterations and what-if analyses, insight is gained from the big data analytics. From the insight, informed decisions can be made. In Step 4, the models and their findings are tested and validated and presented to stakeholders for action. Implementation is a staged approach with feedback loops built in at each stage to minimize risk of failure. The next section describes several reported big data analytics applications in healthcare. We draw on publicly available material from numerous sources, including vendor sites. In this emerging discipline, there is little independent research to cite. These examples are from secondary sources. Nevertheless, they are illustrative of the potential of big data analytics in healthcare.

Table 2 Outline of big data analytics in healthcare Methodology

Step 1 Concept statement

- Establish need for big data analytics

Project in healthcare based on the "4Vs".

Step 2 Proposal

- What is the problem being addressed?
- Why is it important and interesting?
- Why big data analytics approach?
- Background material

Step 3 Methodology

- Propositions
- Variable selection

- Data collection
- ETL and data transformation
- Platform/tool selection
- Conceptual model
- Analytic techniques

-Association, clustering, classification, etc.

- Results & insight

Step 4 Deployment

- Evaluation & validation
- Testing

4. Challenges

At minimum, a big data analytics platform in healthcare must support the key functions necessary for processing the data. The criteria for platform evaluation may include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance. In addition, while most platforms currently available are open source, the typical advantages and limitations of open source platforms apply. To succeed, big data analytics in healthcare needs to be packaged so it is menu driven, user-friendly and transparent. Real-time big data analytics is a key requirement in healthcare. The lag between data collection and processing has to be addressed. The dynamic availability of numerous analytics algorithms, models and methods in a pull-down type of menu is also necessary for large-scale adoption. The important managerial issues of ownership, governance and standards have to be considered. And woven through these issues are those of continuous data acquisition and data cleansing. Health care data is rarely standardized, often fragmented, or generated in legacy IT systems with incompatible formats [6]. This great challenge needs to be addressed as well.

5. Conclusion And Future Research

Effective integration of data mining and medical informatics and its subsequent analysis using big data techniques will no doubt impact healthcare delivery costing and improved healthcare results via well informed decision making (Sun, 2013). From a systematic review of the literature, a cross section of variety of articles was extracted for use within the study. Literature was located in multiple databases, which suggests that yet no standard or definitive natural place for publishing on big data in healthcare has been established. Further, the limited number of academic articles found suggest that as an aspect of healthcare, health informatics and data science, it has yet to establish a sound academic base and body of literature, despite this, there is a plethora of marketing cases and suggested uses evident through the popular press and websites. Future similar research may include analysis of such materials and examples of vendor studies, as well as identification of other material from reference citations from the initial articles selected. From the systematic review of the literature, utilisation categories were created. The categories incorporate semantics. The creation of categories can potentially result in creation of standard terms and vocabulary used within big data analytics. Indeed, this could result in more refined searches and improved yield to support evidence based medicine, and form reliable and efficient decision

support for diagnosis and treatment of patients. Identification of what literature is currently being published regarding big data analytics in healthcare may lead to more defined publishing that can assist in identifying the current uses of big data analytics. In addition, identification of the databases can also result in determining what databases are helpful for future research in big data analytics.

6. References

[1]. Raghupathi W: Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis;2010:211–223.

[2]. Burghard C: Big Data and Analytics Key to Accountable Care Success. IDC Health Insights; 2012.

[3]. Dembosky A: "Data Prescription for Better Healthcare." Financial Times,December 12, 2012, p. 19; 2012. Available from: http://www.ft.com/intl/cms/ s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK.

[4]. Feldman B, Martin EM, Skotnes T: "Big Data in Healthcare Hype and Hope." October 2012. Dr. Bonnie 360; 2012. http://www.west-info.eu/files/big-data-inhealthcare. pdf.

[5]. Fernandes L, O'Connor M, Weaver V: Big data, bigger outcomes. J AHIMA 2012:38–42.

[6]. IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/ wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.

[7]. Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com/ collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf.

[8]. Bian J, Topaloglu U, Yu F, Yu F: Towards Large-scale Twitter Mining for Drugrelated Adverse Events. Maui, Hawaii: SHB; 2012.

[9]. Raghupathi W, Raghupathi V: An Overview of Health Analytics. Working paper; 2013.

[10]. Ikanow: Data Analytics for Healthcare: Creating Understanding from Big Data. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf.

[11]. jStart: "How Big Data Analytics Reduced Medicaid Re-admissions." A jStart Case Study; 2012. http://www-01.ibm.com /software /ebusiness /jstart/ portfolio/ uncMedicaid CaseStudy.pdf.

[12]. Knowledgent: Big Data and Healthcare Payers; 2013. http://knowledgent.com/ mediapage/insights/ whitepaper /482.

[13].Explorys: Unlocking the Power of Big Data to Improve Healthcare for Everyone.https://www.explorys.com/docs/data-sheets/explorys-overview.pdf.

[14].IBM: IBM big data platform for healthcare." Solutions Brief; 2012. http://public. dhe.ibm.com/common/ssi/ecm/en/ims14398usen/IMS14398USEN.PDF.

[15]. Intel: Leveraging Big Data and Analytics in Healthcare and Life Sciences: Enabling Personalized Medicine for High-Quality Care, Better Outcomes; 2012. http://www.intel.com/ content/dam/www/public/us/en/documents/whitepapers/ healthcare-leveraging-big-data-paper.pdf.

[16]. IBM: Data Driven Healthcare Organizations Use Big Data Analytics for Big Gains; 2013. tp://www03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_ for_big_gains.pdf.

[17]. Savage N: Digging for drug facts. Commun ACM 2012, 55(10):11–13.

[18]. Zenger B: "Can Big Data Solve Healthcare's Big Problems?" Health Byte, February 2012; 2012. http://www.equityhealthcare.com/docstor/EH%20Blog%20on%20Analytics.pdf.

[19]. K.Mahalakshmi , Manikandan and Nithyanantham "Unsupervised Learning Technique Using Hybrid Optimization for Non-Functional Requirements Classification" , in International Journal of Advanced Engineering Technology, 2016, pp., 1072-1079.

[20]. Mahalakshmi, K., and R. Prabhakar. "Performance Evaluation of Non Functional Requirements." (2013).

[21]. Mahalakshmi, K., and R. Prabhakar., "Hybrid Optimization Of Svm For Improved Non-Functional Requirements Classification" in International Journal of Applied Engineering Research, 2015, pp.20157-20174.

[22]. M., M., Ansari, F., Dornhöfer Khobreh and M. Fathi, "An ontology-based Recommender System to Support Nursing Education and Training," in LWA 2013, 2013.

[23]. V. Castello, L. Mahajan, E. Flores, M. Gabor, G. Neusch, I. B. Szabó, J. G. Caballero, L. Vettraino, J. M. Luna, C. Blackburn, and F. J. Ramos, "THE SKILL MATCH CHALLENGE. EVIDENCES FROM THE SMART PROJECT," ICERI2014 Proceedings, pp. 1182–1189, 2014.

[24]. R.Keerthika and Dr.C.Nalini. "Image retrieval based on context search mechanism." International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 2, Febraury 2014.

[25]. R.Keerthika and Dr.C.Nalini. "Retrieving datasets with nearest neighbour search using spatial queries." International journal on innovative and research technology, 2014 Volume
1, Issue 10.