iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42636

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Big Data Analytics in Cloud Computing for Scientific Analytics

Ashok Kushwaha[1], Dr. Kalyan Acharya[2]
[1, 2]*Maharishi University of Information Technology, Lucknow*

Abstract: ***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. The 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.***
Keywords: Big data, Analytics, Hadoop, Healthcare, Framework, Methodology.

I. 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 U.S. healthcare system alone reached, in 2011, 150 exabytes. At this rate of growth, big data for U.S. healthcare will soon reach the zettabyte (10^{21} gigabytes) scale and, not long after, the yottabyte (10^{24} gigabytes) [6]. Kaiser Permanente, the California-based health network, which has more than 9 million members, is believed to have between 26.5 and 44 petabytes of potentially rich data from EHRs, including images and annotations.

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, we provide examples of big data analytics in healthcare reported in the literature. Fifth, the challenges are identified. Lastly, we offer conclusions and future directions.

©IJRASET: All Rights are Reserved

II. PRIOR WORK DONE

III. METHODOLOGY

A. 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

B. Architectural 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. As Figure 1 indicates, the complexity begins with the data itself.

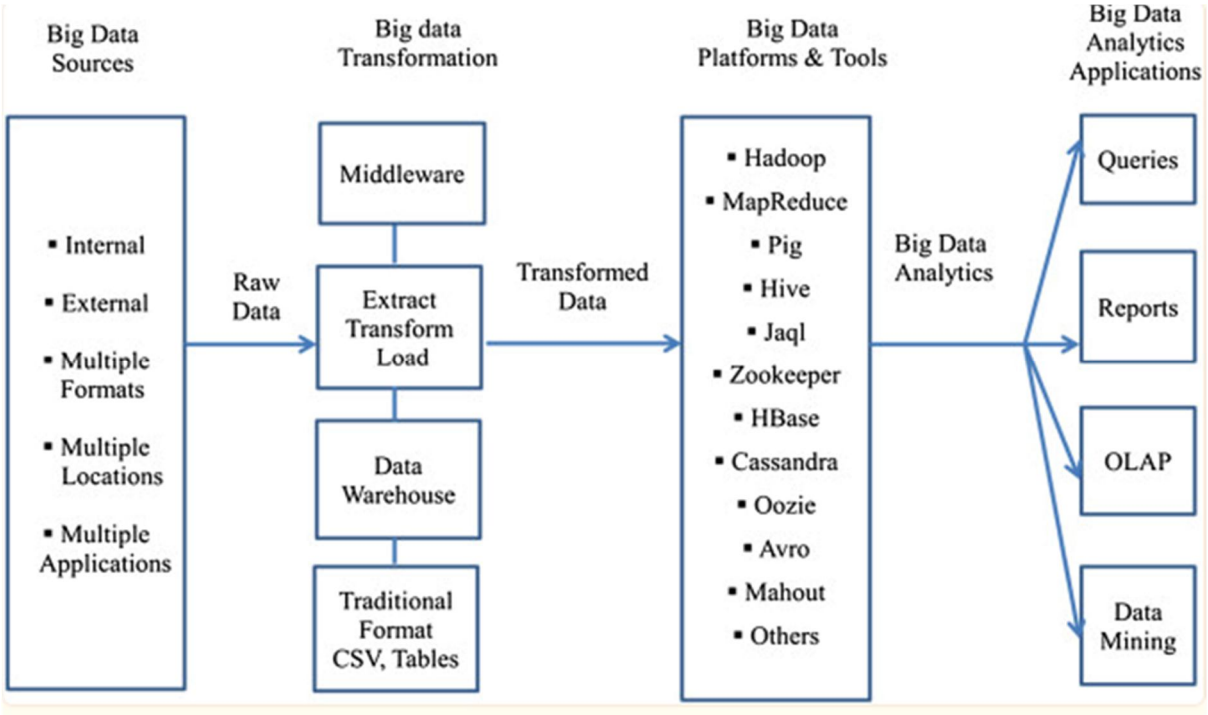


Figure 3.2.1. An applied conceptual architecture of big data analytics

Big data in healthcare can come from internal (e.g., electronic health records, clinical decision support systems, CPOE, etc.) and external sources (government sources, laboratories, pharmacies, insurance companies & HMOs, etc.), often in multiple formats (flat files, .csv, relational tables, 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.

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 has the potential to process extremely large amounts of data mainly by allocating partitioned data sets to numerous servers (nodes), each of which solves different parts of the larger problem and then integrates them for the final result [28–31]. 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, Hadoop makes 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 [30, 31]. These are summarized in Table 1.

Table 3.2.1: 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.
MapReduce	MapReduce provides the interface for the distribution of sub-tasks and the gathering of outputs. When tasks are executed, MapReduce tracks the processing of each server/node.
PIG and PIG Latin (Pig and PigLatin)	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 PigLatin, and the runtime version in which the PigLatin 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 MapReduce 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 (http://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.

Numerous vendors—including AWS, Cloudera, Hortonworks, and MapR Technologies—distribute open-source Hadoop platforms [29]. Many proprietary options are also available, such as IBM's BigInsights. Further, many of these platforms are cloud versions, making them widely available. Cassandra, HBase, and MongoDB, described above, are used widely for the database component. While the available frameworks and tools are mostly open source and wrapped around Hadoop and related platforms, there are numerous trade-offs that developers and users of big data analytics in healthcare must consider. 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.

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.

Table 3.2.2: 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

Source: Adapted from [Raghupathi & Raghupathi, [9]]. 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 1, are also identified; the data is collected, described, and transformed in preparation for analytics. A very important step at this point is platform/tool evaluation and selection. There are several options available, as indicated previously, including AWS Hadoop, Cloudera, and IBM BigInsights. 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.

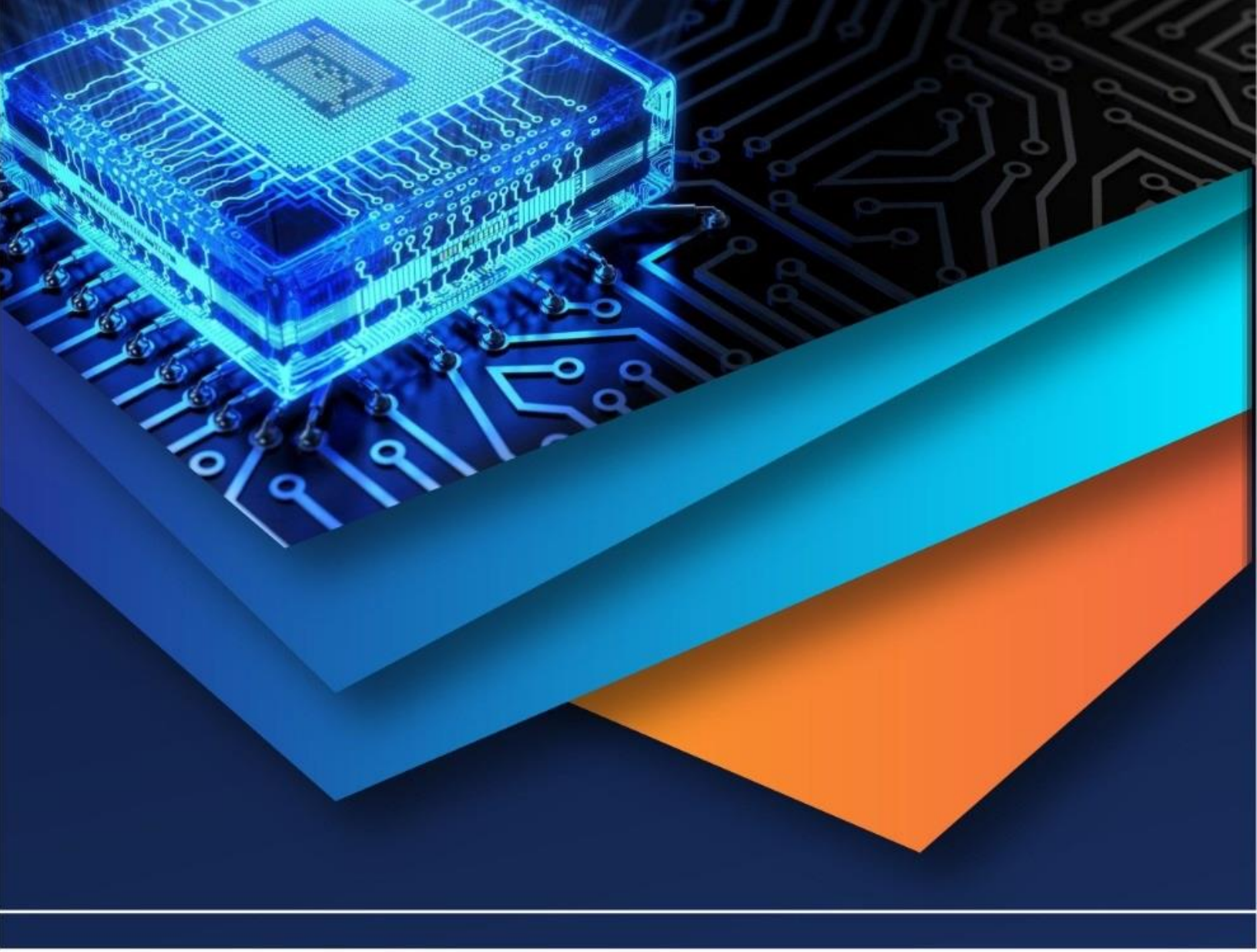
IV. CONCLUSION

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. In the future we'll see the rapid, widespread implementation and use of big data analytics across the healthcare organization and the healthcare industry. To that end, the several challenges highlighted above, must be addressed. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in healthcare are at a nascent stage of development, but rapid advances in platforms and tools can accelerate their maturing process.

REFERENCES

[1] Rocha A., Hauagge C.D., Wainer J. and Goldenstein S., "Automatic fruit and vegetable classification from images", Computers and Electronics in Agriculture, Vol. 70, pp. 96–104, 2010.
[2] Raghupathi W. Data Mining in Health Care. In: Kudyba S, editor. Healthcare Informatics: Improving Efficiency and Productivity. 2010. pp. 211–223. [Google Scholar]
[3] Burghard C. Big Data and Analytics Key to Accountable Care Success. 2012. [Google Scholar]
[4] Dembosky A. "Data Prescription for Better Healthcare." Financial Times, December 12, 2012, p. 19. 2012. [Google Scholar]
[5] Feldman B, Martin EM, Skotnes T. "Big Data in Healthcare Hype and Hope." October 2012. Dr. Bonnie 360. 2012. [Google Scholar]
[6] Fernandes L, O'Connor M, Weaver V. J AHIMA. 2012. Big data, bigger outcomes; pp. 38–42. [PubMed] [Google Scholar]
[7] IHTT . Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry. 2013. [Google Scholar]
[8] 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-technology-complexities-ar.pdf
[9] Bian J, Topaloglu U, Yu F, Yu F. Towards Large-scale Twitter Mining for Drug-related Adverse Events. Maui, Hawaii: SHB; 2012. [PMC free article] [PubMed] [Google Scholar]
[10] Raghupathi W, Raghupathi V. An Overview of Health Analytics. 2013. [Google Scholar]
[11] Ikanow: *Data Analytics for Healthcare: Creating Understanding from Big Data*. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)