ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

AN ANALYSIS OF BIG DATA APPROACHES IN HEALTHCARE SECTOR

Uswa Ali Zia, Dr. Naeem Khan
E-Mail Id: malka_09@yahoo.com
Department of Computer Science, SZABIST Islamabad

Abstract: The big data revolution is in its early days. The majority of the potential for worth creation is still not explored. But it draws the attention of the industry on a track of rapid change which leads to widespread research in this field and contribute to betterment of human life. As digitized medical records are now used by most of the healthcare organization and pharmaceutical companies they have start collecting and storing more and more healthcare data in order to analyse it and obtain insights to solve problems related to variability in healthcare quality , cost , preparedness and safety of healthcare systems etc. This paper provides information about all the significant developments that have carried out so far in the field of big data analysis in healthcare sector. This paper also covers key big data implementation challenges and big data solutions which attempt to solve the challenges of large and fast growing data bulks while reducing the cost and realize its potential analytical value. The paper also has discussed different efforts which have been taken by various researchers to effectively provide solutions using big data analytics to different areas in healthcare sector.

Keywords: Big data, healthcare, Hadoop, MapReduce, sensor informatics, social health, electronic health records, medical images.

1. INTRODUCTION

The term big data refers to the huge amount of data that needs new technologies and architectures to find valuable knowledge from it by using new and innovative analysis practices. Big data is a wide term used for datasets which are large and complex enough so that currently used data processing applications are inadequate for them. In 2012, digital world of data was extended to 2.72 zettabytes which is anticipated to be doubled every two years. Doug Laney expressed the definition of big data as three Vs i.e.

- Volume: the amount of massive data or the quantity of generated and stored data.
- Velocity: denotes the speed of data at which it is produced and managed to fulfil the demands and challenges for its growth and development. Big data is mostly available in real-time. Activities like regular monitoring, e.g. of daily measurements of glucose of a diabetic patient, blood pressure and ECGs.
- Variety: The nature of the data as data comes in all types of formats. This helps people who analyse it to effectively use the resulting insight.

In addition, the complexity referred with data linkage, connection and correlation of the data gathered from different sources and variability i.e. inconsistency of data are also referred as two additional characteristics of big data. The big data works on all types of data i.e. structured, unstructured and semi structured data. The most of its primary focus is to analyse unstructured data. The process of research into huge amounts of data is to reveal unseen patterns and connections named as big data analytics. The amount of data produced and stored globally shows that there is almost a high ability to glean key insights from business information, so far only a small percentage of data is actually analysed. For this reason, big data implementations need to be analysed and executed as much accurately as possible. The importance of big data doesn’t revolved around how much data we haves till, but what actually we want to get from it .So we can take data from any source and analyse it and to find solutions which results in reducing the cost, time, enable new product development and optimized offerings, and making smart decisions.

The area of application of big data includes banking, education, healthcare, government, manufacturers and retails. In the last few years, the big data has been introduced to the healthcare system as providing solution to several of healthcare related information system problems as health systems have grown progressively more multifaceted and expensive. Besides improving profits the big data in healthcare is also being used to forecast or anticipate widespread occurrence of an infectious disease in a community at a specific time frame, cure certain infection, improve quality of life and avoid avoidable deaths. Although it’s a fact that the world’s population is going to be growing rapidly and everybody wishes to live longer and a quality life therefore models of treatment delivery are needed to changing rapidly and many of the decisions behind such changes are data driven. The ambition is to understand about a patient as much early as possible in their life so this can hopefully help in picking up cautionary signs of severe sickness at an enough early stage so that it will help to provide a simple and cheaper treatment. In health division the data collected through mobile devices, caught by health staffs, submitted by entities, or examined in the form of data use, can be a vital tool in considering population health styles. If the data is collected in the perspective of individual electronic health records, this data will help to improve the continuity of care for the individual, but it can also be used to create huge datasets which will helpful in comparing the treatments and outcomes in a more efficient and cost effective way. Numerous platforms and tools are used for big data analytics in healthcare. Certain platforms and tools are available which are used for analysis of big data in healthcare. Some of the common tools and technologies are defined below:

Hadoop: a new technology that permits large data bulks to be processed and organized by keeping the data on the original data storage. Hadoop is open source framework and consist of a Hadoop Distributed File System (HDFS) which allows the basic storage for the Hadoop cluster. It works by dividing the data into little parts and then allocates it over numerous nodes. Hadoop distributed file system assists fast data transfer rates among nodes and allows the system to continue working in a continuous way even if a node get failed. This approach results in lowering the risk of disastrous system failure, even if a majority of nodes become defective. Hadoop is consist of four basic components which includes the libraries and utilities also shared by other Hadoop components often called Hadoop common, Hadoop Distributed File System (HDFS), the programing model named MapReduce and YARN stands for Yet Another Resource Negotiator is a resource management framework with a job of scheduling and managing requests of resource from other distributed applications.

MapReduce: MapReduce a programming model used in Hadoop, which was firstly proposed by Dean and Ghemawat at Google. Hadoop uses MapReduce as the elementary data processing structure. It comprises of two parts i.e. mapping the pieces of tasks on certain slave nodes and then reducing the result from each node into a single result or output. Simply it provides the interface for the sub-tasks distribution and the combining of outputs.

Pig Latin: is a programming language which is used to organize to integrate all types of data (structured/unstructured, etc.). It is comprise of two basic modules: first the language itself, called PigLatin, and second is the runtime account in which the PigLatin code is accomplished. It offers a way to do data extractions, transformations and loading, and elementary analysis without having to compose MapReduce programs.

Hive: it is runtime Hadoop support architecture. It allows SQL programmers to develop Hive Query Language (HQL) statements which are almost similar to classic or traditional SQL statements. Hive helps in analysing the large datasets that are stored in Hadoop's HDFS and well-matched file systems such as Amazon S3 file system.

Jaql: it is a functional, declarative query language that is aimed to process huge data sets. It helps parallel processing by converting high-level queries into low-level queries.

Hbase: this is a column grounded database management system that resides on top of HDFS. HBase tables can function as input and output for MapReduce jobs.

NoSQL: (Not SQL or Not Only SQL) is a database-management system which is not like relational database-management systems, in that they do not use SQL as their query language. They are proved to be better for handling data that doesn't fit easily into tables.

Cassandra: Cassandra is a NoSQL(Not SQL) database which acts as an substitute to Hadoop's Distributed File System.

The most commonly occurring challenges are defined as follow while implementing big data analytics in healthcare industry. The first main challenge is awareness that all of us must fully realize the necessity, complexity and importance of data services. Next is that the powerful companies and enterprises are expected to vigorously promote the implementation of big data technology in the medical industry and drive big data development and talent cultivation. Along with these the big data works on input data in various formats such as documents, e-mails, social media, pictures, videos, sound bites, logs and other forms of information that is difficult to fit into in traditional database tables. Traditional management technologies will improve many of the data volume related issues, but they will not able to solve the non-technical issues associated with data volume properly. Therefore the challenges of big data include scalable data management, data management for large applications, large multitenant databases and large databases security issues for cloud computing, MapReduce and Hadoop environment.

In this report we are critically analysing different approaches used in analysis of big data in healthcare domain and will find out the merits and demerits of these techniques and would suggest a conceptual model to improve the quality of healthcare related information using big data analytics. The rest of this paper is structured into following sections. The Section 2 presents literature review. In Section 3 critical evaluations among various techniques of different papers in tabular form are presented. Section 4 gives future work and Section 5 presents conclusion.

2. LITERATURE REVIEW

Wang et al. [1] draws the attentions of the health care industry on the capabilities of big data and describes some strategies to get full advantage from big data without going deep into the technological view and use it to aid the big data analytics and helps the healthcare organizations respond to the challenges faced by them in a more strategic way in this era of competition. Five main capabilities include traceability, analysis of unstructured data for finding outlines of care, and its foretelling ability to formulate more effective big-data-based strategies by the healthcare managers. The term capability is defined as “the ability to capture, curate, manage, and process the data within a specified elapsed time”. Twenty six healthcare data cases were collected to analyse. Manual content analysis is performed on these datasets to understand the big data benefits and capabilities. Five general categories were documented as a result in which: first, the traceability that means the ability to track data outputted from all the IT systems machineries from all over the organization's service units, to make data visible and accessible for analysis. Second is the unstructured data analytical capability: as most of the clinical data is unstructured so analysis involves data gathering, filtering and visualizing it**.** Only NoSQL (Not SQL) and MapReduce technologies can handle such type of data because the traditional systems were not able to support unstructured data. Third is the analytical capability for patterns of care used to provide a wider view for evidence based medical training by detecting the patterns of care and discovers relations from immense healthcare records. Fourth the decision support capability**:** highlights the ability to generate reports on daily basis. Fifth is the predictive capability which is used to predicting the future outcomes by using statistical analysis or data mining methods. The study also defines five strategies which require further attention of the researchers to visualize suitable big data strategies that will enable healthcare organizations to get fully benefitted from big data in a more efficient way. These includes implementing big data governance, developing the tradition of sharing the information among the employees of the organization, providing analytical training to the staff, presenting cloud computing into the organization and creating new business concepts. The cases analysed in this study shows that big data infrastructure perform as an effective IT item to potentially create IT capabilities and business benefits.

Raghupathiet al. [2] defines the possibilities and potential of big data analytics in healthcare. Along with the potential of big data analytics the study also highlighted several challenges to address. The analysis of big data in the health care sector results in cost reduction and quality treatment to the patients, further benefits includes to identify those individuals who would be benefitted from anticipatory care or by changing their routine in a proactive manner; outlining the broad scale disease to support prevention initiatives; gathering and issuing data on medical actions, identifying, predicting and dropping fraud by applying advanced analytic systems for fraud recognition and checking the correctness and stability of claims. Several challenges are also highlighted which includes governance issues including ownership, security, privacy have however to be addressed. The big data potential is great if the relevant challenges will overcome and by reducing the limitations of the already available open source platforms. By overcoming the existing limitation as defined above will help in more fast progress in analysing the big data in healthcare.

Roskiet al. [3] mainly focus upon adopting the advanced and appropriate IT infrastructure to support the big data analytics in healthcare sector in a more efficient and effective way along with focus to make changes in current polices about data usage, access, sharing and privacy to balance the potential social benefits of big data approaches and the protection of patients' privacy. The IT infrastructure suggested includes the use of data lakes for data provenance, the use of cloud service provider (CSP) for data security and privacy as being more efficient, secure and flexible, and to develop some visualization tools for assistance of data analysis. The study focuses mainly upon adopting the advanced and appropriate IT infrastructure to support the big data analytics in healthcare sector in a more efficient and effective way. Furthermore some policies need to be reviewed to get a balance between data sharing and privacy. The major issue still addressed is the difficulty of protection of the data privacy. The health organizations must identify their goals or the specific problem they need to solve through big data analytics to avoid the expense of investing for implementing big data infrastructure.

Hay et al. [4] tries to maps the geographical areas where there is a greater chance of an infectious disease to be occurred and those areas where the chances are relatively low. The analysis is based upon the environmental factors such as temperature and rain fall. The source data will be gathered from various sources and in various formats need to be processed in real time and thus uses big data techniques to map the surveillance of disease in real time. The input data from various sources is to be processed in real time and uses big data techniques (such as data mining or machine learning) to map the surveillance of disease in real time. Using data mining techniques such as machine learning and the use of multitude sourcing provides an opportunity of creating a continually or frequently updated atlas of infectious diseases. By the use of these easily accessible dynamic infectious disease risk maps would be valued to a vast range of health professionals from policy creators to ranking limited resources to specific clinicians. The study improves the traditional methods of mapping the diseases. Traditional methods provide a continuous risk map in static time. Though using big data analytics techniques it is possible to provide the risk map in real time. The proposed technique will generate the reliable and efficient maps because the novel data sources have issues of reliability so the machine learning process is standardized as using triage process which assigns a weightage to every data point as a degree of reliability. Such weighting is an essential part of mapping techniques to measure the uncertainty of the output from each location. Hence the maps produced using big data techniques are reliable and efficient. Such maps will help the relevant authorities to take accurate predictive measures.

Weber et al. [5] emphasis to identify all the diverse but useful data sources like social media, census records, credit card purchases, and numerous other types of data and then link them together while taking care of the privacy and security, so as to get fully benefitted from big data. As the biomedical data is distributed across different isolated areas so it is necessary to link them all to get better insights from this available data by analysing it. Although before linking data from all sources for analysis it is also necessary to distinguished between the useful sources and the irrelevant data sources. The study applies the probabilistic linkage algorithm for linking the diverse sources. This algorithm's main advantage is that the same technique is used to match the patient's crossways different electronic health records can be stretched to the data sources outside the health care. Although the major challenge to identify the novel data sources and then preserving the security and privacy if the data related to a patient is still present and need attention. The study proposed that one way to solve this issue is to grant all the responsibility to the patient so that it is up to the patients will either to share and decide where to share the data or not.

Ram et al. [6] proposed a model to predict the amount of asthma patients in emergency departments in real time. The study is to get benefitted from twitter, Google search, and environmental air data to evaluate visits for asthma victims in a relatively distinct geographic area in short time duration. The study gathered emergency department visits data related to asthma, collected data from Twitter, internet users' exploration interests from Google, and pollution sensor data taken from the environmental protection agency, for the same geographic zone and same time period in order to create a model for sake of forecasting asthma associated emergency department appointments. The results produced by this model are with 70% precision. This work is different from existing studies that classically predict the spread of transmittable diseases like flu by using social media. As asthma is a non-transmittable health condition and this study shows the effectiveness and value of linking big data from various sources in developing predictive models for such non-transmittable diseases with a particular focus on asthma. This model results will help to provide cautionary signals to the people at high risk for asthma adversarial events as early as possible, by enabling well-timed, pre-emptive, and targeted preventive and beneficial interventions. The study has provided primary evidence that environmental data and social media can be proved helpful to precisely predict asthma emergency department calls at a broader level. As asthma problem is going to rise, new and synchronized strategies must be developed at the public health level and clinical levels to control the social problem of asthma adverse outcomes that will be supported by this model. Moreover, predicted risks could be visualized both location wise and temporally, and should make presented to stakeholders through numerous media sources.

Meredithet al. [7] defines the importance of big data in prevention of certain disease by continually measuring and analysing the data in real time from different sources and suggest precautions to particular individual about his/her disease while lowering the cost. Big data can assist action on the risk factors such as physical activity, nutrition, use of tobacco, and exposure to pollution. These risk factors for disease at are helped at population and individual levels, and by refining the effectiveness of involvements to help people reach healthier behaviours. The study describes two case studies to show how big data is helpful in disease prevention. Disease prevention is based upon to identify modifiable risk factors for disease like exercise, diet, alcohol consumption, smoking and pollution get insights then lead to interventions to improve these risk factors and improve health. The number of mobile health applications is increasing day by day and the data will come from diverse sources helps in improving health behaviours. So it is important to monitor and measure data from these sources as such data is more detailed and modern technological advances have provided many new ways of doing this. First case study explains the big data relation with physical activities, the new devices and smartphone apps that have the potential to passively and continuously track physical activity create a unique source of big data for health. The second case study focuses big data analytics contribution particular to asthma disease on the pattern. The issues of privacy and access to data still need to be solved.

Raoet al. [8] enlightens the security challenges related to big data with particular reference to healthcare sector. As in healthcare sector the security and privacy problems of big data are the foremost alarming as data is bound by certain international regulations like the Health Insurance Portability and Accountability Act (HIPAA). The study aimed to propose feasible security solutions to get fully benefitted from big data relating to healthcare in very controlled environment. The study explains the necessity of big data analysis in healthcare sector to do proactive and reactive analysis of the information which will results in providing chances for forecasting, realizing uncertain needs, and decreasing risks as along with providing tailored services. The issue of patient privacy is although a major concern in the domain of big data analytics. The study describes the challenges of data security into two categories i.e. the challenges related with law and regulations which differs from country to country and secondly the technical challenges which includes the infrastructure and techniques to be used to support big data analysis. The study describes and compares the technical challenges associated with a most popular big data supporting database called NoSQL also known as non SQL or Not only SQL. Though NoSQL is better in performance and highly available and scalable, but including NoSQL all other big data databases should have lack of security features as compared to traditional databases. Big Data platform must hold several layers of security for both the data at rest and in working mode. It should able to encode personal sensitive data inside the data warehouse. Hence there is need to develop some secure solutions for analysis of data. The study also proposed four security models which are data de-identification model, data centric approach to security, walled garden model and jujutsu security. The study proposed some requirements which should be preliminary to every anticipated security solution. These includes limiting the access of both users and applications to the original data, encryption of the sensitive data, securing data with fast ,efficient and low cost solutions, implementing block layer encryption, use of user friendly security tools, security should also be implemented into the technology. The most important aspect of the security solutions is that they should compatible with both the legacy system and new technologies. On the basis of these requirements the paper proposed four types of security models which are data de-identification model, data centric approach to security, walled garden model and jujutsu security model. Healthcare sector use big data analytics to convert data into actionable information by creating data-driven visions to intelligent business and clinical decisions like dropping admissions rate, recognizing and removing waste, enhanced clinician workflow etc. Healthcare data is really sensitive, with privacy and integrity as a main attribute. Therefore security is vital in healthcare’s related big data. Security solutions should be implemented in such a way that they should guarantee safe analytics and securing big data frameworks. The future work includes developing some better security models or modifying the proposed models to develop a balance between the basic requirements that should be present in the security models.

Augustine et al. [9] focuses on the benefits of using Hadoop for the analysis of big medical data (images) produce in healthcare sectors. The study emphases on the benefits of using Hadoop for analysing big medical data as being more flexible, scalable and as a more economical solution. The study aims to analyses the images produces in healthcare sectors. Hadoop provides solution to analyse the medical images by combining these medical images from numerous sources and extracts the important data for accurate diagnosis. For this purpose the study emphasis on the use of an interface called Hadoop Image Processing Interface (HIPI) supports the image processing as accomplished in Hadoop. Hadoop Image Processing Interface works in a distributed fashion. The Hadoop Image Processing Interface takes the input in the form of HipiImageBundle (HIB). Then a culling function is applied on these images to assess them on some predefined criteria and reject those images which fail to meet the desired criteria. A function namedCullMapper class is then applied on every single image which passes the culling test. The Images are inputted as FloatImage to the Cullmapper class with an associated ImageHeader. This process will be cheap because Hadoop uses industry standard hardware therefore the cost per terabyte of storage is comparatively ten times inexpensive than a traditional data warehouse. The big data analytics is in early stages of gain its place in healthcare sector in India. The study focuses that by giving little more attention to Medical Image Processing will results in reducing the cost of services to a common man in the country. The problem is that still in India most of the patient’s data is not digitized hence not easily available and accessible for building any type of statistics and analysis. So the need of hour is to digitize the medical records and introduce consistency in storing medical records by various Acts.

Srinivasan et al. [10] detects fraud and abuses in the healthcare insurance claims by developing two big data analytics applications named as CMC- (Capital Markets Cooperative) I+PLUS and CMC-HIBIS (Health Insurance Business Intelligence Services). The I+PLUS uses predictive modelling techniques, while HIBIS uses business intelligence techniques. These applications help to identify the hidden patterns that results in cost overrun. The proposed techniques discover anomalies in the insurance claims. These applications specially provide assistance to private health insurance companies. These two proposed applications provide effective analytics as well as logical explanations for alerts that will aid to accelerate or fasten to take appropriate action. The applications works on large bulks of complex organized and text free data dig out from both the insurance statements and hospital discharge data to discover the claiming outlines which signify fraud, abuse, and faults. The main contribution includes the comparative analysis of current and past data obtained for specific time frame when a patient was in a hospital. Many similar solutions were also available but they havelack of operational interface and will not able to compare the past and current data to find the anomalies. One of the proposed application (CMC-I+PLUS) uses predictive modelling technique for analysis. It identifies the anomalies and generating reports on abnormalities related to cost and quality of care. While the second application (CMC- HIBIS) uses business intelligence to detect and report fraud through alerts. The most vital element of CMC-HIBIS is the alert explorer which serves as a visualization tool and helps the insurer staff to take appropriate actions on the bases of the type of alert generated. The Alert Explorer is a vital component of CMC-HIBIS which serves as the visualization tool. It provides the facility of search, manage, view and review the details. The claim anomalies detected as a result of these applications will help of which the private health insurers improve an unseen cost overrun which was not easily detected by previous or traditional transaction processing systems.

Salian et al. [11] explains that analysing the big data will help in predicting the risk of diabetic patient’s readmission efficiently by determining the risk predictors that can be a reason of readmission of diabetic patients. The study suggested a predictive model that can find the patients with chronic diabetes diseases and are most likely to be get admitted again and again. In the suggested system works by loading the raw data is loaded into the Hadoop File System (HDFS) firstly and then by using Hive queries, all the nominated predictive variables are recovered into a comprehensible dataset to use for modelling. And then model works by selecting and applying various classifications, prediction method using Hadoop. The accurateness of the results was checked by confusion matrix. The top five readmission predictors in diabetic dataset are body mass index, plasma glucose, age, pregnant, pedigree function are top predictors in the proposed model. This study shows that the risk of readmission for diabetes patients can be evaluated by big data analytics. Predictive modelling has been worked by applying decision tree classification method. The chance of readmission in diabetic patient is successfully predicted by this proposed model. The results show that the feature plasma glucose is extremely linked with readmission feature whereas that feature blood pressure is least associated with readmission.

Eswariet al. [12] uses the prediction model by using analysis algorithm in Hadoop/Map Reduce environment to predict the widespread diabetes types and the related complications and also the treatment. The suggested architecture of predictive analysis system is constructed on various levels e.g. data collection, warehousing, predictive analysis, processing analysed reports. The system analysis by working in Hadoop/Map Reduce environment to categorize the type of diabetics, its problems and the type of treatment suggested for such patients. The final results are then dispersed over many servers and simulated through several nodes subjected on the geographical area. The information of individual patients will be now be exchanged among health care centres that will leads to provide proper treatment at correct time in distant locations at low cost. The suggested system uses Hadoop as the open-source distributed data processing platform. Hadoop has the ability to play both roles of data organizer as well as analytics tool. Big Data Analytics in Hadoop's implementation provides organized way for attaining better results like availability and affordability of healthcare service to population as this research ambitions to deals with the study of diabetic's treatment in healthcare industry using the big data analytics.

Gowsalya et al. [13] aims to propose a system with the ability of predicting the risk of readmission of diabetic patients within coming 30 days by measuring the probability with help of MapReduce technique. This risk factor obtained will aids the physicians in suggesting suitable care for the patients. The study presents solution which uses Hadoop MapReduce to analyse huge datasets and mine useful observations from the dataset that helps in assigning the resources effectively. For new patients, this system makes use of the information of the prior patients with similar conditions and reuses those suggestions. The system collects the data directly from the patients (body sensors) and their corresponding doctors. This data is then stored on Hadoop Distributed File System (HDFS) and MapReduce technique is applied by HDFS. Analysis is performed on the datasets with information of hospital admission, diabetic encounter, laboratory tests, medications, length of stay in the hospital. The rate of the readmission is calculated on the features like age, HbAIC result and modification in prescription. Haemoglobin A I C (HbAIC) is an considered important factor as a measure of glucose control, which is mostly results to measure of diabetes. The likelihood of getting readmitted is high if the value is greater than 8%.The use of distributed file system for the development of this proposed system uses low cost present hardware and stores data across nodes. This predictive system helps hospitals and other health care organizations to assign clinicians, nurses, machinery, and other resources in a better way. Patterns in the patient admissions, duration of stay, and other factors can be analysed and measure and used to anticipate future volumes mostly at extreme times. It can improve the outcome of patient treatments and help the providers in making more intelligent judgments about treatment, by reducing the complications and specially reducing the chances of the patients to get readmitted in the hospitals.

Sadhana et al. [14] aims to analyse the huge diabetic's data sets to reveal some interesting fact which will help in developing a prediction model. The diabetes becomes a global hazard and will increasing rapidly. The study emphasis on the need to analyse the already available huge diabetic data sets to analysed so to discover some vital facts which may help in producing some prediction model. Besides using the data mining techniques (as previously used) this study is going to uses Hadoop, hive and R for analysing the datasets. The datasets were taken from Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. Total eight attributes (no. of pregnancies, glucose plasma concentration, blood pressure, serum insulin, body mass index, age, diabetes pedigree and skin fold depth) were used to produce the result as a patient being effected by diabetes when output is 1 and not when indicated 0. The raw csv file is injected to hive as input where these datasets were analysed on the basis of these attributes. The output of hive is given to R as input which performs statistical analyses along with producing the graphs. The basic benefit of Hive is that it acts as a data warehousing solution that constructed on top of Hadoop. It offers a query language similar to SQL termed as HiveQL. It is used for querying tables which are stored in form of flat files on Hadoop Distributed File system (HDFS) with whole Meta data repository and also cares data portioning on a specific factor. The results produced are highly efficient as hive has analysed 768 datasets in just 19 seconds. The graphs generated by R can help to understand the outcomes in a simpler manner. The graphs are generated for all eight attributes and explain the effect of these attributes on chance of someone to be getting diabetic. The study claims that a prediction model should be developed by using such graphs or information.

Sharmila et al. [15] aims to analyse the big data in predicting the diabetes from medical record of the patients. From last few years, the number of people suffering from diabetes gets increased and it was reported as one of the major cause of deaths because of its complications which will results in harm of kidneys, heart and nervous system. The study states that approximately 40 million Indians suffer from diabetes till now. This number is expected to double by 2020. Therefore it is important to predictor diagnose the diabetics in a patient as early as possible. This study is analysing the diabetes from huge medical records by using decision trees with statistical implication using R.R is a sequential programming language for the analysis, graphics and software development activities for data mining and in various fields. About lakh of datasets were collected from Chennai to analyse having ten attributes (i.e. pregnant, LDL, post prandial HDL, BMI, HBAIC, age, creatinine, family) and a class variable. There are four possible outcomes i.e. either the patient is positive for diabetes (shown by output one), prediabetes (shown by output 2) gestational diabetes (if output is 3) non-diabetic (shown by output 4). The csv file are loaded into R. after the prepressing the decision tree algorithm is applied to predict all the four possible diabetics outcomes as defined above and produces the results. The R tool analyses a lakh datasets in just 748.54 seconds. Moreover the correlation coefficient for two attributes was calculated after the data set is analysed using R. The obtained value of correlation coefficient is 0.4, which shows perfect positive linear relationship between variables. This study uses R tool which is quite effective, extensible and having comprehensive environment for statistical computing and graphics. Another important feature of R is that it supports a variety of file formats (XML, binary files, CSV) and also user created R packages. The study also uses decision trees for the reason that they are easy to understand, economical to construct, easy to incorporate with database system and is relatively accurate in several applications. In this study a thorough analysis of the diabetic datasets was done efficiently with the help of R. this information which was discovered from this study can be further used to develop efficient prediction models.

3. CRITICAL EVALUATION

Based on the literature review reported in the previous section, a critical evaluation of the different big data analytics approaches is provided in Table 1.

Table-3.1 Critical Analysis of the Big Data Analytics Techniques

Ref	Focus Area	Proposed Techniques/ Solution	Strengths	Limitation / Weaknesses	Possible Improvements	Validation parameters
[1]	Elaborating big data capabilities	The strategies are proposed to get full advantage from big data by ignoring technological view and to aid the big data analytics.	These capabilities of big data achieve improvements in healthcare sector while reducing the cost and enable the managers to develop long term strategies to get benefitted from the huge volume of data.	The study merely focused on analysis of unstructured data.	Healthcare industries need to seek effective IT artefacts, and to develop models based on these strategies.	Effectiveness
[2]	Possibilities and potential of big data analytics in healthcare	The study proposed benefits and methodology, describes examples stated and then concisely deliberates the challenges and then offers conclusions.	Certain techniques and technologies described along with their advantages and limitations, also comparing the emerging technologies for big data analysis with the traditional techniques and technologies of analysis.	The study is limited to highlight the benefits of big data analytics in the healthcare sector.	To built the user friendly big data analytical tools and to resolve managerial issues of governance, ownership and standards.	Efficiency
[3]	The focus is mainly upon adopting the advanced and appropriate IT infrastructure to support the big data analysis in healthcare sector in a	The study proposed IT infrastructure relevant to the issue of data security, data privacy and integration.	The study proposed IT infrastructure required for big data's storage sharing access, and integration. And by reviewing some policies as pointed by the authors, the healthcare organizations will get better insights from this data.	The study is restricted to looking into policy implications with regard to adopting big data analytics in healthcare sector.	The healthcare organizations and their policy makers have to change their old mind sets and hold new approaches while overwhelming the barriers to data sharing with suitable	Effectiveness and Efficiency

	more effective and efficient manner.				protection and mutually working towards the goal of delivering better health outcomes at low cost.	
[4]	To maps the geographical areas where a greater chance of an infectious disease to be occur.	The input data from various sources is to be processed in real time and uses big data techniques (such as data mining or machine learning) to map the surveillance of disease in real time.	The whole atlas of existing distributions will be of significant benefit to increase upcoming calculations of the problem related to the severity of occurrence of a disease and helps in taking precautionary measures accordingly.	The major problem is the limited participation of target audience particularlythe public and private R&D.	The practical implementation of advanced risk mapping approaches using big data analytics techniques will enable to produce maps for different types of infections, at different scales, and for different purposes, such as risk levels.	Reliability and efficiency
[5]	To identify all the diverse but useful data sources and then link them together while taking care of the privacy and security.	Probabilistic linkage algorithms are used.	Linking the diverse data sources.	Merely implemented on high risk patients as because it requires expensive infrastructure and complex algorithms for analysis. The study is incapable to provide any effective standards to be implemented to retain the privacy of the data.	A mechanism to retain patient's data privacy need to be developed while sharing data	Usability and accessibility.
[6]	To calculate the amount of asthma patients in emergency departments in real time.	The asthma predictive model be dependent on on a mixture of electronic medical records and by analysing emergency department visits, Twitter tweets Google data and sensor data. A combination of classification techniques, natural language processing techniques and machine learning are used to develop this model.	Results can be supportive for public health observation, emergency department awareness or readiness and targeted patient interventions.	Limited to English language tweets and to restrict to emergency department visits data of a single hospital only. Furthermore the noisy data is not handled properly.	To discover the effect of related data from other types of social media e.g. discussion forums, and blogs on this model. And extending this model to the diseases with temporal and geographical variability such as diabetes.	Accuracy and reliability.
[7]	Role of big	Focused on analysing	The diverse data	Limited to	Prevention	Effective

	data in disease prevention	data from different sources especially mobile sources in real time.	sources are more detailed and big data tools and techniques will enable to get insights from this data in real time and improve public health at low cost and efficient way.	explain the sources of data to analyse and to help in disease prevention.	model should be developed that will be cost effective and should accessible on individual level.	ness
[8]	The security challenges related to big data with particular reference to healthcare sector.	Proposed four security models (de-identification model, data centric approach to security, walled garden and jujutsu model)based on some predefined requirements of security.	Each model has its own strengths related to security of data at different levels.	Does not provide any practical evidence to the measure the success of using any of these models in the real environment.	To develop better security models that balances between the security requirements and the limitations of these four models.	Security
[9]	The benefits of using Hadoop for the analysis of big medical data (images) produce in healthcare sectors.	The interface called Hadoop Image Processing Interface (HIPI) supports the image processing as accomplished in Hadoop.	The application uses MapReduce technique which increases the efficiency while executed on large data.	Limited to the analysis of big data medical images.	Similar techniques should be developed or used for analysing the audio and video.	Effective ness
[10]	To detect fraud and abuses in the healthcare insurance claims.	Two big data analytics application named CMC- (Capital Markets Cooperative)I+PLUSan d CMC-HIBIS (Health Insurance Business Intelligence Services) were proposed. The I+PLUS uses predictive modelling techniques, while HIBIS uses business intelligence techniques.	The applications provide complete claims based intelligence to inspect the potential irregularities in health claims. HIBIS enables the claims processing and acquiescence staff to examine the claims and reject a claim. The alert explorer of HIBIS serves as strong visualization tool.	The applications can only be helpful for ordinary medium sized insurer and serve only as nationwide applications.	Social network analysis methods will be used to study provider member associations and behaviours, and by data mining techniques to explore useful information from descriptions regarding cure and procedures.	Reliabilit y and usability
[11]	To find the risk predictors and then a detailed analysis has been done to predict risk of readmission of diabetic patients on basis of these	The suggested system practise builds a predictive model which identifies the patients with diabetes diseases and are most likely to be get readmitted. By using Hadoop MapReduce, Hive and R language and Hadoop File System (HDFS).	Such analytic methods help in suitable consumption of resources in hospitals and reduce the cost incurred due to re-hospitalization.	The prediction model only works for diabetic patients.	Such type of predictive models should be developed for other chronic diseases.	Accuracy .

	predictors.					
[12]	To anticipate the widespread diabetes types, difficulties related with it and the treatment to be provided with the help of predictive analysis.	Predictive analysis algorithm used in Hadoop/Map Reduce environment to predict the diabetes types (type 1,type 2,type 3) prevalent, problems linked with it and the cure to be provided.	This system is the podium for knowledge and intelligence based prediction in real time management of huge size of data.	Doesnot provide any validation parameters to check the accuracy of the results produced.	A proper set of parameters need to be explored for correlation among multiple data sources available on the net. Such type of attribute correlation can help to prepare a better treatment plan for diabetic patients.	Effectiveness
[13]	Predict the risk of diabetic patients to be admitted again in the hospitals in upcoming thirty days.	The proposed solution uses Hadoop MapReduce to analyse huge datasets and mine valuable insights from the dataset which helps doctors to allot resources effectively.	The algorithm yields a 'risk score' lies in between 0 to 3 (low to high) for every admitted patient by using MapReduce techniques.	The study is limited to the diabetic patients with limited prediction of thirty days.	Future work includes to find patients with alike prior circumstances and same readmission rate. They are then grouped on the test results and treatments. This may result in improving collective diagnosis of chronic diseases.	Usability, reliability
[14]	To analyse the huge diabetic datasets to discover some interesting facts and develop a prediction model on these basis.	The study analyses the large datasets by using Hadoop, Hive and R. The simple raw datasets were injected to Hive and hive performed analyses on these datasets and then the output of Hive is given as an input to R which produces graphs.	Hive supports a partitioning the table on a particular factor. Secondly Hive disclosures the best portions of Hadoop, i.e. Map Reduce and data storage, to users who do not knows about map reduce or not concerned for creating Map Reduce programs.	The system only works on limited set of attributes.	This analyses can be further utilized to develop some effective and efficient prediction model	Usability, visibility
[15]	To analyse big medical records and predict diabetes on early stages.	The analyses should be done by using R tool and decision trees which will produces the accurate output in just few seconds.	The analyses are done by using R tool because it is extensible, effective, having strong visualization capability and can work on several file formats. Similarly the decision trees are used as they are easy to understand, cheap to construct, easy to assimilate with database system and are more accurate.	The study is only restricted to the analyses of massive amount of medical records.	This information can be further used to develop efficient prediction models. Future work will also consist of parallelization using multiple cores can also be used to improve the prediction model using R.	Accuracy and effectiveness

CONCLUSION AND FUTURE WORK

The major gaps found from section II and section III includes lack of predictive models, low accuracy rates, security and privacy issues and lack of user friendly tools available to deal with big data analytics. Therefore it will be concluded that Big Data is neither a problem nor a solution, it is basically an opportunity. This study aims at exploring this opportunity within the healthcare domain. Big data analysis in healthcare systems allows to leverage unbelievable amounts of data and processing resources related to healthcare. Healthcare systems need to allocate more time and resources for planning and forecasting. The recommendations for healthcare organizations to get benefitted from big data analysis includes establish centre for business intelligence that will put the entire focus on big data and then achieve their business goals in shorter time span. This will be accomplished by having better understanding of different big data analysis techniques, technologies and their implementation along with security, internal and external system integration, hosting and development platforms. Moreover, it is observed that there is much more potential to generate some predictive models for diabetic datasets to help overcome the complications of this vastly widespread disease.

REFERENCES

[1] Y. Wang, L. Kung, C. Ting and T. Byrd, "Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care", 2015 48th Hawaii International Conference on System Sciences, 2015.

[2] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2, no. 1, p. 3, 2014.

[3] J. Roski, G. Bo-Linn and T. Andrews, "Creating Value In Health Care Through Big Data: Opportunities And Policy Implications", Health Affairs, vol. 33, no. 7, pp. 1115-1122, 2014.

[4] S. Hay, D. George, C. Moyes and J. Brownstein, "Big Data Opportunities for Global Infectious Disease Surveillance", PLoS Med, vol. 10, no. 4, p. e1001413, 2013.

[5] G. Weber, K. Mandl and I. Kohane, "Finding the Missing Link for Big Biomedical Data", JAMA, 2014.

[6] S. Ram, W. Zhang, M. Williams and Y. Pengetnze, "Predicting Asthma-Related Emergency Department Visits Using Big Data", IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1216-1223, 2015.

[7] M. Barrett, O. Humblet, R. Hiatt and N. Adler, "Big Data and Disease Prevention: From Quantified Self to Quantified Communities", Big Data, vol. 1, no. 3, pp. 168-175, 2013.

[8] S. Rao, S. Suma and M. Sunitha, "Security Solutions for Big Data Analytics in Healthcare", 2015 Second International Conference on Advances in Computing and Communication Engineering, 2015.

[9] D. Peter Augustine, "Leveraging big data Analytics and Hadoop in developing India’s healthcare services," International Journal of Computer Applications, vol. 89, no. 16, pp. 44–50, Mar. 2014.

[10] U. Srinivasan and B. Arunasalam, "Leveraging Big Data Analytics to Reduce Healthcare costs”, Costs", IT Professional, vol. 15, no. 6, pp. 21-28, 2013.

[11] S.Salian and G. Harisekaran, “Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients,” International Journal of Science and Research, vol. 4, April 2015.

[12] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive methodology for diabetic dataanalysis in big data," Procedia Computer Science, vol. 50, pp. 203–208, 2015.

[13] M. Gowsalya, K. Krushitha, and C. Valliyammai, "Predicting the risk of readmission of diabetic patientsusing MapReduce," pp. 297–301

[14] S. Sadhana and S. Savitha, “Analysis of Diabetic Data Set Using Hive and R,” International Journal of Emerging Technology and Advanced Engineering, vol. 4, July 2014.

[15] K. Sharmila and S. Manickam, “Efficient Prediction and Classification of Diabetic Patients from bigdata using R,”International Journal of Advanced Engineering Research and Science, vol. 2, Sep 2015.