A Survey on Applications of Big Data Analytics in Healthcare

Shubham Borikar, Mohan Bhagchandani, Raunak Kochar, Ketansing Pardeshi, Manisha Gahirwal

Abstract*—The data in healthcare is increasing rapidly and is expected to increase significantly in coming years. Healthcare services although armed with modern technologies for curing the diseases grapples when it comes to preventing the diseases beforehand. Adoption of Big Data solutions will play an important role in transforming the outcomes of the healthcare industry by promoting evidence based reasoning and providing patient centric treatment. In this age of Big Data we can provide solutions to identify individuals who are prone to certain lifestyle diseases. Think of identifying an individual having an increased risk of diabetes after 10 years, now. With the advent of new big data analysis tools and technologies, such predictive systems can be designed which can identify individuals with increased risk. This paper provides an overview of big data analytics, different technologies that can be used in big data and its impact on healthcare domain to make some useful predictions based upon analyzing a variety of datasets. Finally we provide a model which can be used for predictive analytics using data mining and machine learning algorithms to predict the chances for a person to be prone to a disease.

***Keywords*—Big Data;Healthcare;Prediction system**

I. INTRODUCTION

The healthcare industry has been generating data in large amounts. Traditionally these records were being kept in written form however the current trend has been towards digitizing these records. The data in the healthcare sector is growing rapidly and is coming from various internal as well as external sources like mobile devices, wearable sensor devices, clinical notes, social media etc. The data that is generated is in petabytes which cannot be processed by relational databases efficiently. Also data required for relational databases is structured while big data techniques can process structured as well as unstructured data.

By efficient incorporation of Big Data in healthcare we can effectively create model which would provide an informed view of health data. This would ameliorate the decision making process where the biological knowledge appears to be restricted. Effective analysis of the present health data can help in providing newer solutions to the present diseases. Properly examined, the data can also provide vital statistics which can provide information on epidemic outbreaks, disease risk prediction, public policy drafting, etc. Thus healthcare data can help us delve deeper in developing preventive systems which would in turn provide economic stability to all sections of the society.

Big Data refers to collection of data sets considered to be too large and complex to easily manage and process using traditional database management tools. The main descriptors of big data are defined as 3 Vs: Volume, Velocity and Variety. Volume refers to a lot of data which cannot be efficiently processed by traditional ways. Big Data sizes usually include data sets with sizes from a few terabytes to many petabytes. Velocity refers to the data that accumulates swiftly. Big data is usually composed of collected observations over time, and/or recorded transactional data points, for a distinct set of entities. Variety means that the data comes from a lot of different places. That means that there is structured and unstructured data from different sources involved, such as different systems, different databases and different applications [8].

II. TOOLS AND TECHNOLOGIES

A wide range of tools and products are available in big data analytics like open source Apache Hadoop based analytics, Stream computing software for real time analysis and Data warehouse for operational insights. HDFS is Hadoop Distributed File System that enables storage of large files by distributing data among pools of data nodes.

MapReduce is a programming model that allows for massive job execution scalability against cluster of servers.

Hive is a data warehouse system based on MapReduce programming model and is built on top of Hadoop. It makes use of high level SQL like statements and its performance is much better than My-SQL on traditional databases. PIG is Perl-like language used for query execution over data stored in Hadoop cluster. Apache Storm can be used for stream processing of information like social media feeds. R is a language for statistical computation with very powerful graphics capabilities and works very well with Hadoop. It is free and open source software. Rapid Miner provides template based framework to do advanced data analytics. It is an advanced analytics platform that can execute in-memory, in-cloud, in-database, in-stream and in-Hadoop. Weka is a set of machine learning algorithms for solving data mining related problems. It supports many of the Data Mining processes such as Preprocessing, Clustering, Classification, Regression and Visualization [4].

Revised Version Manuscript Received on October 30, 2015.

Shubham Borikar, Bachelor of Engineering, Student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Mohan Bhagchandani, Bachelor of Engineering Student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Raunak Kochar, Bachelor of Engineering Student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Ketansing Pardeshi, Bachelor of Engineering Student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Manisha Gahirwal, Professor, Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

III. APPLICATIONS OF BIG DATA ANALYTICS IN HEALTHCARE

The big data solutions can be used in the health care to get innovative outcomes in the following areas:

- Personalized healthcare - Predictive data analysis systems can provide early detection of a disease before a patient actually develops disease symptoms. Pattern detection through stream mining from real time wearable sensors for elderly or disabled patients can be done to alert the physicians if there are any changes in vital health parameters.
- Secondary usage of health data - Deals with agglomeration of clinical data from government, patient care, administrative records to discover valuable insights like identification of patients with specific disease, therapy choices, clinical performance measurement etc.
- Drafting public policy - Big data solutions can aptly provide tangible summarized data basis for effective drafting of the public policy.
- Population health - Analytics solutions can mine web-based media data to predict future trends.
- Evidence based medicine- Evidence-based medicine involves the use of quantified research and statistical studies by doctors to form diagnosis. This enables doctors to make better decisions not only based on their own judgement and perceptions but also from the best available evidences. It also provides a means of validating and verifying scientific hypotheses with statistical health models [4].

IV. CHALLENGES

Some of the challenges of Big Data in healthcare industry are:

- No fixed standards for health care data - Unlike other fields, primarily, there is no established or mutually accepted data aggregation standards across the healthcare industry throughout the world. There is a vast amount of healthcare data that is generated by different agents in health care today, ranging from insurance claims to general practitioner notes, data about health in social media, and streaming data from wearable sensors and other health monitoring devices.
- Integration of heterogeneous data models - With the advent of electronic health records there is emerging problem of interaction between legacy and modern systems. Heterogeneous data from different sources like electronic health records (EHRs), hospital systems, labs, et cetera fragmentation is difficult to merge into an integrated standardized database system.
- Infrastructure Issues - Hospitals already have a Legacy system and their compatibility with new technologies always remain an issue. Such conflict can be moderated using middleware systems to convert all the data to appropriate models before processing
- Insufficient real time processing - Time delay in processing continuous streaming data models could lead to less quality patient care.
- Data Quality - Incorrect data can lead to misleading, incorrect or useless information, which if pertains to healthcare data can be dangerous too. In order.to get reliable insights from the data for making patients health care related decisions, the quality of the data is very important [4].

V. CURRENT SYSTEMS

1. Care Architecture:

The architecture of CARE system can be seen in Figure. The basic steps for the algorithm are given below.

- In the first step an individual inputs a set of diseases. The set is the accumulation of diseases over their medical history.
- The individual's diseases are then compared to all other patients available in the existing database and an initial filtering is done.
- With this filtering only those patients with whom individual has some disease similarity are kept.
- On this filtered dataset collaborative filtering is performed.
- Final output is a list of diseases which the patient can have [5].

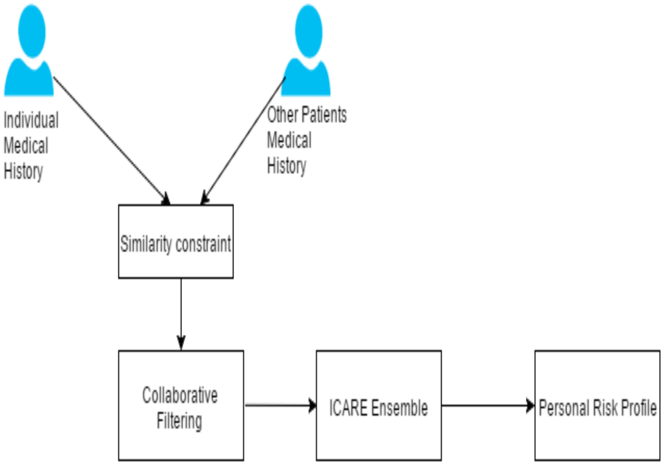

Figure 1-Care Architecture

2. Three tier architecture:

The tree tier architectural model is used to collect heterogeneous data from different sources, converting it to a standard form, analyze it and provide valuable insights.

- Data collection - Heterogeneous health care data is collected from different sources.
- Data extraction - The data that is extracted from multiple sources and stored on a single NoSql database.
- The extracted data is converted into a standard form.
- Data analysis - Using various analytical methods and technologies such as data mining algorithms, in-memory computing etc. analysis on the data is done to gain valuable insights.
- Data Interpretation - Proper interpretation of the result by an expert with clinical support is important as improper interpretation can convey different meaning [4].

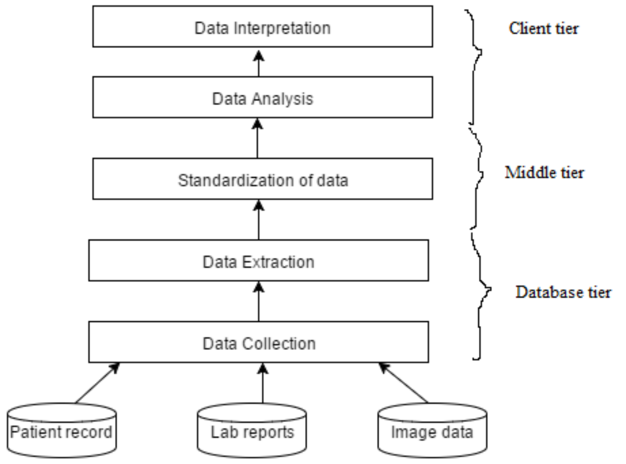

Figure 2-Three tier architecture

VI. PROPOSED SYSTEM

Predictive system in Healthcare is a system which basically takes various health parameters of a user as an input and provides a comprehensive report on the risk of user towards various lifestyle diseases. The system primarily functions in two stages; firstly, it performs descriptive analytics on various available datasets to identify the vital parameters which drive the development of particular diseases; secondly, it performs predictive analytics using user submitted data and performing regression analysis techniques. The entire system would be developed in R Statistical Analysis language using open-source IDE R-Studio.

In the highest level of abstraction, the system primarily performs big data analytics on user data and provides a comprehensive health report which incorporates the risk value of the user towards contracting various lifestyle diseases. The system would apply various clustering algorithms in order to identify vitals. The entire system would be implemented using open-source software. In future the system would try to implement stream mining techniques to provide real time health analysis by incorporating various wearables, data collectors, etc. Various lifestyle diseases like Type 2 Diabetes, Arteriosclerosis, etc. are generally characterized by one or more health factors such as High BMI, large waist size, high glycerides etc. However, these types of diseases can be follow a pattern and are influenced by the above mentioned factors and these factors influence the diseases in varying proportions. Also these diseases can be controlled if appropriate health care measures are taken. Our system would assist the individuals to assess their health conditions now, and take preventive actions so as to prevent such kinds of diseases.

Our system can have varied applications and can be implemented across the domains. The primary aim of the system would be to serve individual users. Assessing individual health and providing effective health report card to caution the individual against future diseases. It would also assist the medical specialists to delve deeper into identifying the causes and the interrelations of the diseases.

The prominent aims of our proposed system are as follows:

- Devising a healthcare prediction system which would incorporate a wide range of lifestyle diseases.
- Deciding the prominent factors influencing the risk of contracting the diseases in the near future by variable subset selection.
- Identifying various interrelated diseases based on clustering techniques.
- Creating different dashboards for various applications and different users like individual users, doctors, government agencies and insurance companies.
- Scalability to incorporate additional disease identification after additional data agglomeration.
- Providing healthcare suggestions to the users. Generating a report to provide personalized health tips to the users of the system.

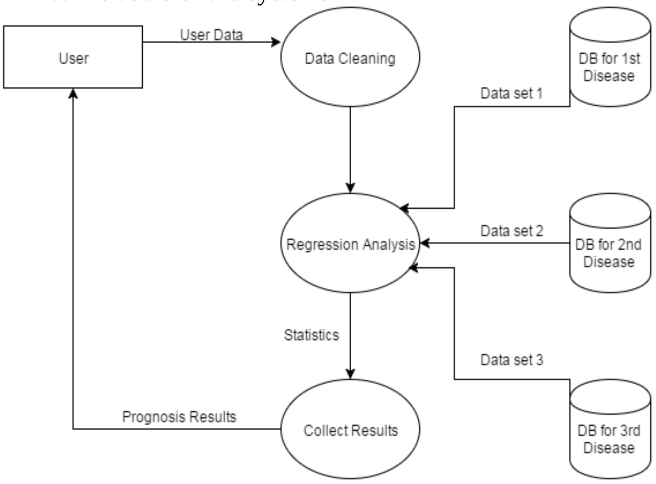

Figure 3-Proposed System

VII. CONCLUSION

Though there are several challenges like combining heterogeneous data, infrastructure issues, insufficient real time processing, data quality that must be addressed, Big Data has the potential to transform and revolutionize the way healthcare systems use technologies to gain valuable insight from the data repositories. In the future we are sure to see widespread use of big data analytics across the different areas of healthcare industry. This paper provides various big data tools whose proper selection can give promising results. Big data analytics and its applications in healthcare are at an initial stage of development, but rapid advances in its platform and techniques can accelerate their maturing process.

ACKNOWLEDGMENT

We would like to take this opportunity to thank our mentor and project guide, Mrs. Manisha Gahirwal, Professor of the Department of Computer Engineering at V.E.S.I.T., for her continued support and guidance.

REFERENCES

1. Nambiar, R.; Bhardwaj, R.; Sethi, A.; Vargheese, R., "A look at challenges and opportunities of Big Data analytics in healthcare," in Big Data, 2013 IEEE International Conference on , vol., no., pp.17-22, 6-9 Oct. 2013

2. Wullianallur Raghupathi, Viju Raghupathi “Big Data Analytics in Healthcare: Promise and Potential” http://www.hissjournal. Com/content/2/1/3 Health Information Science and Systems 2014, 2:3
3. Viceconti, M.; Hunter, P.; Hose, R., "Big Data, Big Knowledge: Big Data for Personalized Healthcare," in Biomedical and Health Informatics, IEEE Journal of , vol.19, no.4, pp.1209-1215, July 2015
4. Mathew, P.S.; Pillai, A.S., "Big Data solutions in Healthcare: Problems and perspectives," in Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on , vol., no., pp.1-6, 19-20 March 2015
5. Keith Feldman, Nitesh V. Chawla, "SCALING PERSONALIZED HEALTHCARE WITH BIG DATA", 2nd International Conference on Big Data and Analytics in Healthcare, Singapore 2014.
6. J.Senthil Kumar, N.Ramprasath, "A Scrutiny on Current and Parallel Big Data Analytics in Health Care", International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume 12 Issue 4 –FEBRUARY 2015.
7. Kaul, C.; Kaul, A.; Verma, S., "Comparitive study on healthcare prediction systems using big data," in Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on , vol., no., pp.1-7, 19-20 March 2015
8. Durham, E.-E.A.; Rosen, A.; Harrison, R.W., "Optimization of relational database usage involving Big Data a model architecture for Big Data applications," in Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on , vol., no., pp.454-462, 9-12 Dec. 2014
9. AbuKhousa, E.; Campbell, P., "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems," in Innovations in Information Technology (IIT), 2012 International Conference on , vol., no., pp.267-272, 18-20 March 2012

AUTHORS PROFILE

Shubham Borikar is a Fourth Year Bachelor of Engineering student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Mohan Bhagchandani is a Fourth Year Bachelor of Engineering student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Raunak Kochar is a Fourth Year Bachelor of Engineering student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Ketansing Pardeshi is a Fourth Year Bachelor of Engineering student at Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.

Manisha Gahirwal is a Professor at Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, University of Mumbai, Mumbai (Maharashtra). India.