International Journal of Research in Advent Technology, Vol.7, No.3, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

A Survey on Big Data Tools and Its Use in Predictive Analysis in HealthCare

Jay L Borade[1], Joel Dsouza[2], Gunjan Munde[3], Divya Varghese[4]
Information Technology[1, 2, 3, 4], Fr. Conceicao Rodrigues College of Engineering, Bandra[1, 2, 3, 4]
Email: jay.borade@fragnel.edu.in [1], joel.reujoe@gmail.com[2], mundegunjan96@gmail.com[3], ddivya.varghesee@gmail.com[4]

Abstract-Current trends and statistics show that technologies like Big Data and Machine Learning algorithms together used can change the way we process and analyze data. Both technologies can be said to be booming in a variety of industries. There are many aspects of Big Data and predictive algorithms based on Machine Learning that are recognized and being considered by researchers and technology consultants around the globe. Inspection of Big Data in healthcare shows a high level of opportunities that can be used to determine causal relationships between different healthcare components. This analysis can provide valuable relationships that can be used in IT transformation, practices and business value. In this paper we attempt to get an insight on different research that have been conducted by collaborating technologies like big data, machine learning and electronic health records. This collaboration can be used to bring many benefits in the field of healthcare. We breakdown the discussion into four sections beginning with the introduction, methodology, results and conclusion

Index Terms-Big Data; HealthCare; EHR;.

1. INTRODUCTION

Data processing is one of the greatest inspirations in the areas of research with the fundamental focal point of dissecting gigantic measures of information. In the present time the blend of big data, machine learning, cloud computing and information warehousing are getting to be famous in the field of social insurance. Healthcare data as Electronic Health Records (EHR) can be made, utilized, put away and recover essential patient information. The volume of EHR is excessively to be appropriately utilized on any scale. Any information identified with human services, for example, physician notes, biological information, lab reports, patient metadata, case history, diet routine and rundown of specialists in a specific clinic can be considered as healthcare data **[1]**. We have to comprehend what kind of data we are managing and what noteworthy job does it play in procuring learning. Data can be briefly be thought of as tacit or explicit **[2]**. Tacit learning can be said to be information created because of individual experience. This sort of learning can be extremely abstract and could be exceptionally unpredictable. Then again explicit learning is the point at which it simple to gather organize and convey information of different people. With regards to healthcare we can consider distinctive things like therapeutic reports and records that give assistance in getting explicit knowledge. It could likewise incorporate restorative techniques and diagnosis of specific diseases. Diverse data sources can be viewed as with regards to medicinal services investigation like EHR, Genetic Data, Medical Imaging, Documentation and Test Reports. Big Data in healthcare is concerned with the how it provides and adheres to the 5 V's of Big Data i.e. Volume, Velocity, Variety, Veracity and Value. Talking about the first V, Volume is the amount of data coming in. A proper process has to be defined that can ensure that volume does not affect the way we achieve outcomes. Velocity is the rate at which data is generated and moves around. Variety speaks about what different source we get our data from. Veracity refers to the worthiness or truthfulness of data that we obtain and finally we speak about Value i.e. does the data provide any value that we need. There might be a clear link between data and insight it does not however mean that value is provided by Big Data. After we are clear with the concept of how Big Data needs to adhere to the 5 V's we can describe what useful analysis can be obtained by the use of predictive algorithms in Machine Learning that can in turn provide some predictive analysis about the data we obtain. We consider algorithms like CNN-UDRP and CNN-MDRP that can be used to predict diseases in a geographical region

2. METHODOLOGY

This section briefly gives an overview of different tools that can be used in big data and healthcare analysis.

2.1. Tools For Big Data

Hadoop Architecture:
Apache Hadoop utilizes the ace slave design where in you have two hubs specifically the datanode and name-node. The name-node executes as an ace and data-node as a slave **[3]**. The name-node deals with the entrance to data-nodes. The data-nodes then again administrate and store information over numerous hub

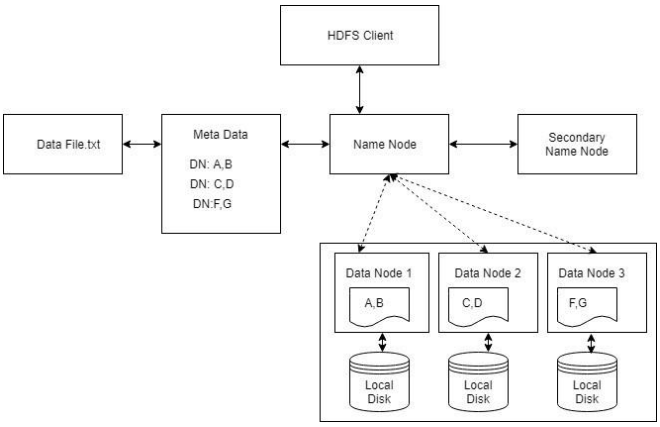

Fig 1. Hadoop File System Architecture

Hadoop MapReduce:
The Hadoop MapReduce is a programming model used to process huge size of data sets across a Hadoop Cluster **[3]**. Hadoop framework also provides the scheduling, distribution, and parallelization services to process the big data **[3]**.

Comparative analysis of different big data frameworks:

Hive: It has a SQL-like (HiveQL) interface implemented in MapReduce used for structured data. It is used as data warehousing application. The Thrift server in Hive allows external clients to interact with Hive over a network, similar to the JDBC or ODBC protocols

HBase: It is mainly used to modify data in an HDFS environment and is useful as a distributed database. Serves Data-driven websites like Facebook (up until 2018) to implement messaging platform

Pig: It contains the SQL like (Pig Latin) language used for semi/unstructured data and is widely used for data analytics. It is capable to fit user-code at any junction of the pipeline

Sqoop: It helps mainly to pull the data into the Hadoop Big data platform and is considerable option for data management. It has the ability to import and export bulk of data from databases

Spark: It works as real time data processing engine and is used for processing data in memory. It is capable of processing huge data amounts like facebook or twitter. It is elastic and can be used for faster deployment.

2.2. Data Warehouse Based EHR platforms

OpenEHR and EHR4CR are the most widely recognized platforms utilized as data distribution center for EHR examination. The EHR4CR utilize something known as a Common Information Model (CIM). A CIM is a higher request question connected on EHR information distribution center. EHR information is unstructured information as indicated by CIM. We have to apply the ETL activities so as to get unstructured information into organized information and the EHR4CR deals with nearby data models and the EHR4CR CIM. OpenEHR is structured dependent on prerequisite caught through numerous years. The base prerequisites need to construct OpenEHR framework are, information distribution center incorporates EHR, model stores, wording, and statistic or character data **[3]**. The statistic vault is utilized as a front end to store persistent ace record (PMI). The EHR can be arranged to incorporate either no statistic or some recognizing information **[3]**. Plainly the above stages are a decent alternative for the investigation of verifiable human services information. Anyway in the present time the driving needs will in general execute new innovation other than information warehousing so as to deal with the new prerequisites.

2.3. EHR, Big Data Systems and Retrieval of EHR records

An aggregate methodology of organized and unstructured information coming from clinical and nonclinical methods of presence will assist us with understanding and foresee illnesses. Figure 2 demonstrates a fundamental system of more entangled stage which can process big data of EHR and give us more attractive execution in confounded analytics as opposed to data warehouse platform. Truth be told, this technique enables healthcare organizations to gainfully record an aggregate clinical showdown quickly and recoup essential information from EHR big data cluster in high execution preparing. Along these lines, this methodology gives most requests that are required in huge EHR era. To look at tasks and to deliver understanding that empower decision makers and to make strides and upgrade health performance and utilitarian effect, Healthcare establishments made data warehouses. An augmentation in information ought to urge the Healthcare associations to move to big data technology to give new highlights that are referenced in data warehouse approach. Parallel and distributed computing are a portion of the fundamental models that are utilized in the handling of big data, due to which it can execute process simultaneously on number of machines. An open-source bundle called Hadoop was released by Apache for distributed data handling with as of late patterns that can be used in long term for healthcare. Hadoop Distributed File System otherwise called HDFS can access, handle, and recover all information records at the same time among the Hadoop group.

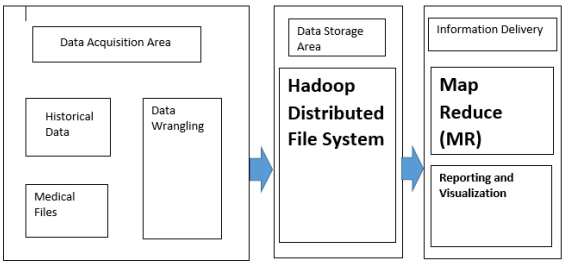

Fig-2. Big Data Analytics in Healthcare

In **[4]** it is talked about that how big data frameworks can be utilized for structuring stages for directing elastic search. So as to lead elastic search we can creep numerous sites. Concentrate information from these web structures and file them by elastic search techniques. After the whole system is directed a customer can look by setting a search key and search alternative. The means of leading elastic search is additionally quickly expounded in [4] as below:
Stage 1: Collection and capacity
Stage 2: Classification
Stage 3: Create Index
Stage 4: Search and Analyze
A few uses of big data and elastic search can consequently be utilized so as to create manageable applications for simplicity of search useful patterns that can be utilized in long haul for healthcare

2.4. Predictive Analysis of Big Data

When it comes to chronic diseases it is observed that nearly 50% of the American population suffer from some chronic disease and nearly 80% of healthcare fee is spent on this treatment **[5]**. If only there was a way to provide an early identification of a chronic disease based on patients information we could come up with ways to improve our lifestyle. Say we have some information about a patient such as age, gender, living habits and we could output some cerebral output which indicates whether the patient is in a high risk or low risk of the population **[5]**. Here we can speak of mainly two algorithms used for disease prediction namely CNN Unimodal Disease Risk Prediction (CNN-UDRP) and CNN Multimodal Disease Risk Prediction (CNN-MDRP). CNN-UDRP can be described as 5 stage process where first we represent data as text using convolutional filters in the form of dimensional word vectors and then apply it to a convolutional layer of text based CNN that can extract a feature. This feature then works as an input to the pool layer. The pooling layer takes care of selecting only the text that play an important role. The output of the pool layer is then provided to a full functional CNN which in turn provides input to the classifier for classifying a particular disease.

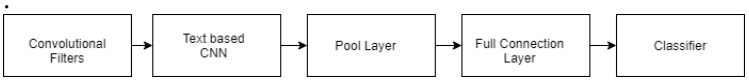

Fig-3 CNN-UDRP

CNN- MDRP on the other hand is based on training the algorithm on data that can be structured as well as unstructured and training the data that is set on a word vector dimension of 50 or more. Both algorithms can be used in identifying if a geographical region has a risk of incurring any chronic diseases. This is particularly useful for insurance companies that want to estimate whether people belonging to a certain age group have a potential to garner a disease based on which a policy can be decided. Along with MapReduce tools that allow for extraction and arrangement of data the above mention algorithms can be useful when both are combined thus obtaining the best possible outcome.

3. SOME APPLICATIONS FROM DISCUSSED PAPERS

On perception we see that there exists numerous innovation and instruments that can be utilized so as to create frameworks that can give simple access to social insurance information. On joining the Big Data like Hadoop and OpenEHR or EHR4CR we can create frameworks where simple access to EHR is conceivable. Following are a portion of the outcomes that can be gotten on the blend of the above devices

(1) Healthcare examination: On investigation a patients past reports and current wellbeing checks a specialist could show signs of improvement acumen on giving medicine
(2) Prognosis: Prognosis is only the early conclusion of interminable age related illnesses.
(3) Report Management: Quick access to ones reports and wellbeing data any specialist can think about what treatment should be given to a patient
(4) Healthcare Service Recommendations: Better administrations could be given by examination of Healthcare Service Providers based on client evaluations and criticism
(5) Follow up Systems: Follow up Systems could be utilized so as to monitor a patients current wellbeing status concerning his/her last visit.

Hence the previously mentioned applications can be conveyed to reality by the mix of two basic accessible innovation stacks

4. CONCLUSION

These days, space (from healthcare institutions to home and carry) and time (from distinct sampling to consistent tracking and checking) are not anymore a hindrance for modern healthcare by utilizing better

and stronger analysis techniques. Medical diagnosis is evolving to patient centric expectation, and treatment. The big data techniques have been created throughout the years and will be executed all over. Subsequently, healthcare will likewise enter the big data period. Well basically, due to big data advances, it can be utilized as guide in way of life, as an apparatus to help in decision making, and as a source of development in the developing healthcare system. This paper has exhibited a smart healthcare system helped by cloud and big data, which incorporates 1) a brought together information accumulation layer for the combination of open therapeutic assets and individual wellbeing gadgets, 2) a big data enabled and information driven stage for multisource heterogeneous health care data storage and investigation, and 3) a bundled API for developers and a unified interface for clients. Empowered by Health-CPS, distinctive tweaked applications and organizations are made to address the troubles in the customary human administrations, including united resources, information islands, and patient uninvolved participation. Later on, we will concentrate on creating different applications dependent on the Health-CPS to give a superior situation to people.

REFERENCES

[1] Gunasekaran Manogaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas and Revathi Sundarsekar, "Big Data Knowledge System in Healthcare" Springer International Publishing AG 2017 C. Bhatt et al. (eds.), Internet of Things and Big Data Technologies for Next Generation Healthcare, Studies in Big Data 23, DOI 10.1007/978-3-319-49736-5_7

[2] Komal Sindhi, Dilay Parmar and Pankaj Gandhi, "A Study on Benefits of Big Data for Healthcare Sector of India" Springer Nature Singapore Pte Ltd. 2019 D. K. Mishra et al. (eds.), Data Science and Big Data Analytics, Lecture Notes on Data Engineering and Communications Technologies16,https://doi.org/10.1007/978-981-10-7641-1_20

[3] Youssef M.Essa, Gamal ATTIYA2, Ayman El-Sayed and Ahmed ElMahalawy, "Data processing platforms for electronic health records" Received: 15 February 2017 /Accepted: 3 January 2018 IUPESM and Springer-Verlag GmbH Germany, part of Springer Nature 2018

[4] Aiqin Yang, Shiwei Zhu, Xianyi Li, Junfeng Yu, Moji Wi and Chen Li, "The research of Policy Big Data Retrieval and Analysis based on Elastic Search" 978-1-5386-6987-7/18/$31.00 ©2018 IEEE

[5] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang* Disease Prediction by Machine Learning over Big Data from Healthcare CommunitiesAuthor Biographical Statements DOI 10.1109/ACCESS.2017.2694446, IEEE Access

(A.1)