Trends and techniques of handling torrents of big health data towards enhancing Healthy life Span

Dr M V BRAMHANANDA REDDY,PROFESSOR & PRINCIPAL

DEPT OF COMPUTER SCIENCE AND ENGINEERING

SAI RAJESWARI INSTITUT OF TECHNOLOGY, LINGAPURAM ,PRODDATUR.

Email Id: bramhareddysrit2019@gmail.com

Abstract:

The current trend of society generates torrents of data across various sources like social networking, health sectors, mobile sensors, industries. This voluminous data raised a scope for uncovering hidden insights of this data. This huge data often called big data could undergo several data analytics to retrieve the unnoticed patterns, trends, associations, querying, and information security. Here, in this paper we focus on health care industry towards applying analytics on the health data like EHR's, medical images, reports, sensors and transform this data to make out a meaningful outcome that helps towards diagnosis and prognosis at an early intervention which reduces the morbidity, sensitizing the adverse effects of infectious diseases[2]. We also discuss the existing mechanisms of handling health care data and its underlying effects that are to be tackled.

Introduction:

In the day to day society Internet plays a prominent role in the schedule of every individual for one or the other activity like data gathering, browsing, knowledge extraction, learning, communication, coming to communication activity the social networking sites like Face book, Twitter, LinkedIn alone generate massive amounts of data for a given day. This data raised starting from gigabytes(GB), to Terabytes(TB), Petabytes(PB), Zetabytes(ZB), Yottabytes(YB) and ranging towards Exabytes.

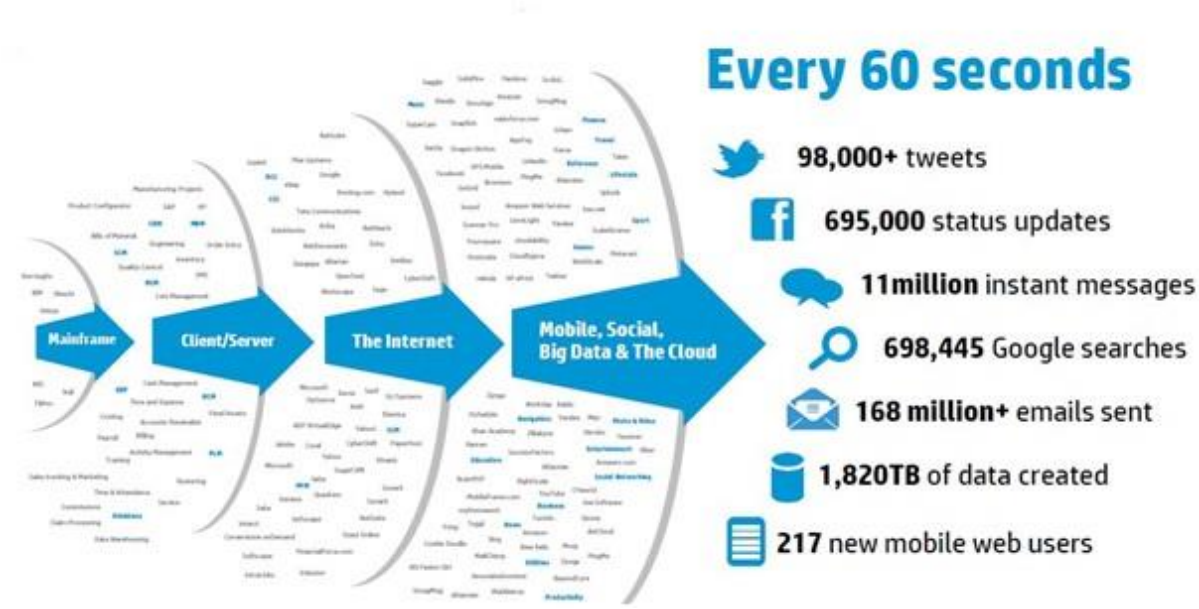

FIG 1: RELATED TO STORAGE CAPACITY

This data is not just one particular form but contains a mixture of contents like text, documents, images, audio, video etc. Storing such huge amounts of data is a major issue. Later comes the concept of information retrieval i.e. gaining meaningful insights from such voluminous data often termed as big data. Any data cannot be called Big data to be called so, it need to satisfy the basic three V's stated according to the META GROUP in 2001.The basic V's of Big data are Volume, Velocity, Variety. Volume refers to the huge amounts of data collected across different sources, Velocity refers to the speed at which the data is generated, and variety refers to the various categories of data present in the collected one. Apart from this another two V's like veracity and value are also added where veracity refers to the quality of data i.e. maintained and value refers to the originality of data.

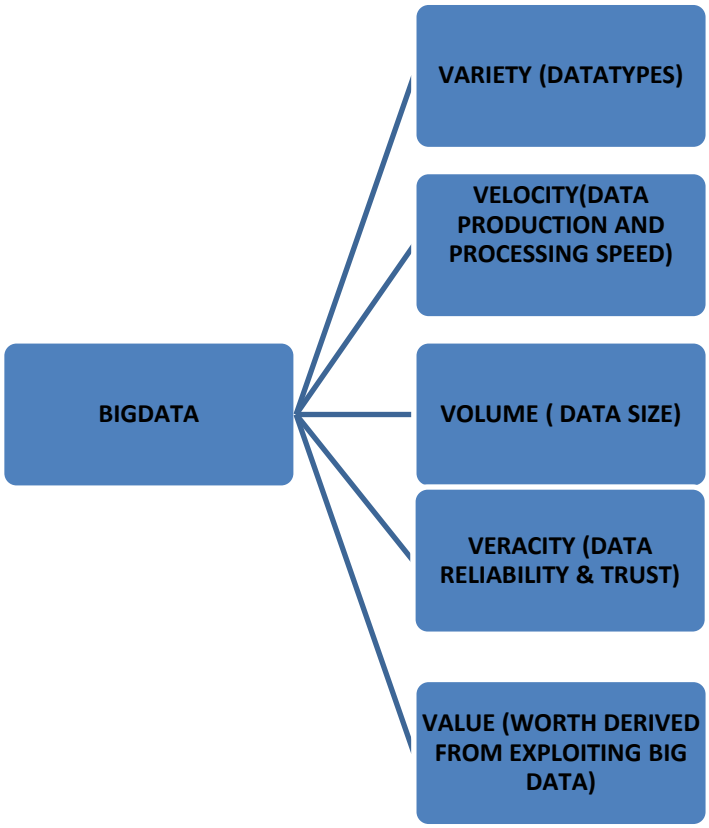

FIGURE 2: REPRESENTING THE BASIC V'S OF BIG DATA

Now a day's people spend much on two aspects of life they are luxury and health. Luxury may denote shopping, entertainment, assets, share market, business sector etc. To make a meaningful outcome of this data we generally prefer a type of analytics called Behavioural analytics, Business analytics, sentimental analytics, market analytics, risk analytics and so on. Behavioural Analytics refers to scenarios where we often see advertisements of our interested sectors during our access of web pages. This is made possible through behavioural analytics where certain features like age, gender can be considered as factors of filtering. Business analytics may refer to drawing associations of several relevant products like bread, butter, milk and eggs stored across could enhance business level, profits in a super market. Sentimental analytics may be useful to study the emotions of people collecting the smiley's, emotions shared across chats and predict the mental status of a person based on some survey.

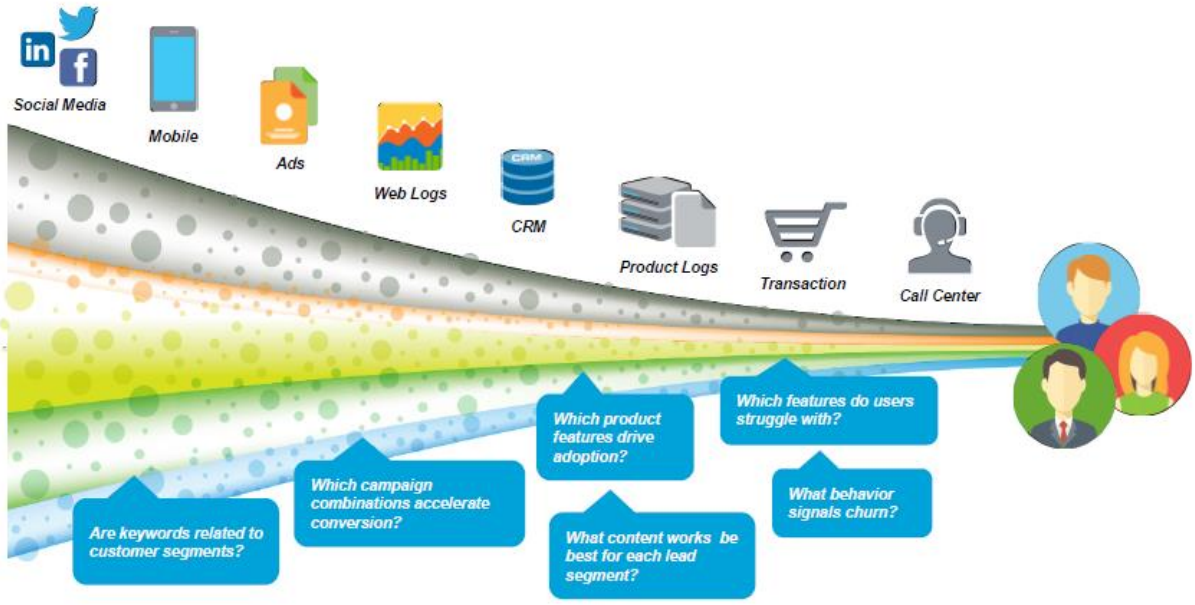

FIGURE 3 : BUSINESS OF BIG DATA

Market Analysis may refer to value of shares, trade, business etc...Risk analysis refers to predicting the level of risk indulged in the specific activity handled. Another area that requires to be focused these days is the healthcare sector. Many organisations offer several health policies, insurance to their employees and even a common man concentrates to attain several insurance schemes, towards safeguarding one self. To offer any services to its end user the health care providers gather the information of individuals from hospitals, surveys, WHO(world Health Organisation) and census data.

The health care data collected is often stored as an EHR(Electronic Health Record), which contains the minimum details of an individual like his height, weight, name, age, gender, BMP, Blood Pressure, Sugar levels etc..This health data collected across different sources may be EHR's, medical images, reports, sensor data and so on[18].

The complex environment of BIG Data consists of multiple data sources along with sophisticated analytics and multiple output forms. It consists of Data Integration, Data Management, Data Analytics and Decision Management. Deployment options are multiple for Big data and user can access through multiple devices.

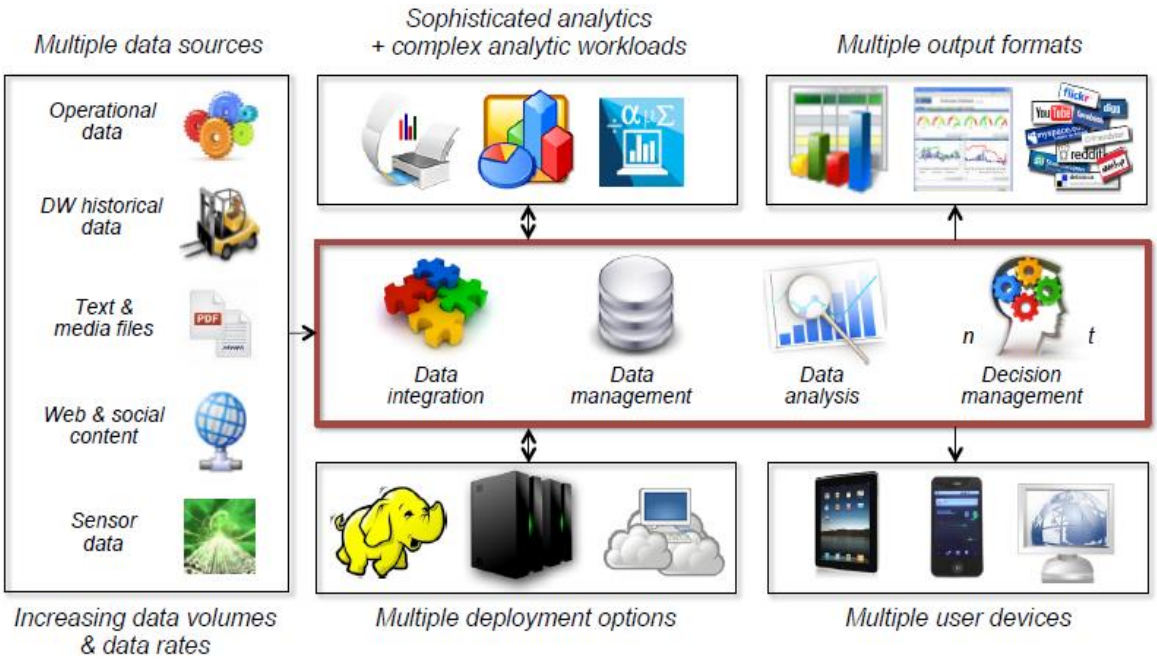

FIGURE 4: COMPLEX ENVIRONMENT OF BIG DATA

Investigative Computing for Health care Industry has Hadoop as an Example.

"If Hadoop didn't exist we would still have to make decisions about what can come into our data warehouse or the electronic medical record (and what cannot). Now we can bring

Everything into Hadoop, regardless of data format or speed of ingests.

This paper focuses towards Architecture of Big data, challenges, applications of big data in health industry, transforming this healthcare data to make necessary outcome, Predictive analysis, Diagnosis & prognosis, Tools and techniques used.

I. Architecture of Big data:

Handling of Big data is a challenging task since, data is generated in massive amounts at a very high velocity which is not sustainable by the traditional databases thus we look out for a new approach of handling this data, as we have not only structured data as in traditional databases but most of the real time data is semi structured or unstructured. The first task starts up with collection of data across different sources then comes the Data Storage thereafter retrieval of data to mine the necessary information to draw valuable insights. There are several platforms where Big data can be handled. In general we use a platform called Hadoop, since it's an open source, end user can program the necessary code. Hadoop contains a HDFS (Hadoop Distributed File System) which concentrates on distributed storage and fault tolerance of the data. On top of this, we use Map Reduce technique which maps similar patterns, clusters sets of data and store them[16] [17]. In the data processing we apply many machine learning algorithms to analyse the data sets collected across and stored in the data base. Not only proper storage of the data but retrieval of required information within the

specified time is a challenging task where we concentrate on Map and reduce algorithms for data accuracy and to gain value of the data

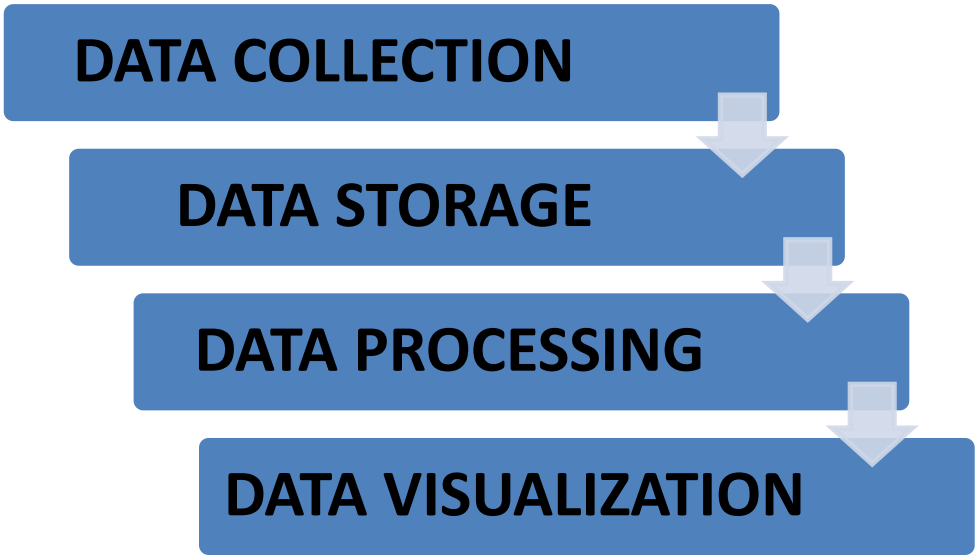

FIGURE 5: ARCHITECTURE OF BIG DATA

Health data volume is expected for a drastic growth in the coming years. Health insurance policies, reimbursement models, various life insurance policies are emerging trends of today's environment. Apart from this the data coming from various providers, hospitals stored as Electronic Health Records (EHR), Electronic Medical Records (EMR), Personal health Records (PHR) provides scope for digitizing the data, store and retrieve it to meet the purpose of single to multi physicians, individuals, providers etc..When big data is synthesized and analysed the aforementioned patterns, associations and trends can be revealed. The application of bigdata analytics in healthcare has potential benefits like early intervention, evidence based medicine, prevention and optimal management that helps in enhancing individual and population health by precautionary measures drawn using predictive analytics

II. Transforming Big health data:

There are various steps involved in transforming big data to analytics:

1. Data Collection
2. Data Storage
3. Data Processing
4. Data Visualization

   The data collection process involves collecting healthcare units of data across various available sources. This data has high heterogeneity since each hospital or organization follows up a different set of approach in projecting their data. In general the healthcare data gathered can be used for health surveillance, predictive modelling, early interventions, optimal disease management & appropriate treatment (Prognostics). Google Flu Trends has collected keywords often used in the search

engine to collect insight on population affected by Flu or Influenza using text analysis considering various features like keywords, geographic area, age and so on.

The reason behind emerging volumes of health data is in US its mandatory to to store the patient data as per the 2009 ARRA (American Recovery and Reinvestment) act. A section of ARRA is called the HITECH (Health Information Technology for Technical Health) act In digitization of health data most of the health data is often collected as individual reports often called Electronic Health Records (EHR). Also depending on the health status of an individual various other reports like diagnostics, medical images could add on. An advantage of maintaining health data as EHR is it could be easily stored and circulated across doctor for treatment, provider for insurance policies, reimbursement and individual for verifying his current health status with the previous instead of physically carrying across the reports. Security aspect can also be added to this EHR restricting its access to concerned doctors, providers and individual by providing an authenticated login[6].

2. **Data Storage**: Previously the data was handled by traditional databases which follows a structured format on which querying and retrieval of data is a smooth process. But now, most of the collected data either exists in Semi structured or Unstructured format i.e data is flat file of no prescribed format, Managing such massive data with Traditional DB is a challenging task since scaling with traditional database is very expensive, Big data cannot fit into size of a single computer. Therefore we choose other tools like the Hadoop which has HDFS file system, MongoDB which supports unstructured data, NOSQL technologies, PIG, Hive etc. Also further the storage of this massive amounts of data could be interrelated to the cloud storage for easy convenience of the end user[19].

3. **Data Processing**: Before processing the data to draw the necessary insights pre-processing of data is a crucial activity before integrating, cleaning, filtering the data. Once done the prepared data is estimated for a model where we use several analytics tools. There are several Traditional analytics tools used in statistical programming like IBM SPSS, SAS, STATA, R language which is an open source, Matlab. The big data analytics offers different categories of analytics like Descriptive or Fact Analysis which denotes what happened, Diagnostic Analytics which denotes why it had happened, Predictive Analytics which describes about what is likely to happen and Prescriptive analytics which explains about what we can do about it[7].Ex: If we consider a student record of a particular university from which we need to know, how many failures, distinctions, why they had failed, how many could clear, what would we do about enhancing it i.e. suggestions and recommendations[3].

In healthcare sector we concentrate mostly on diagnostics, predictive and prognostic approaches of analytics where we need to analyse the root cause of any particular disease, determine what is likely to happen about the disease[8] with respect to the current symptoms and what would be prescribed treatment that could be followed up[13].

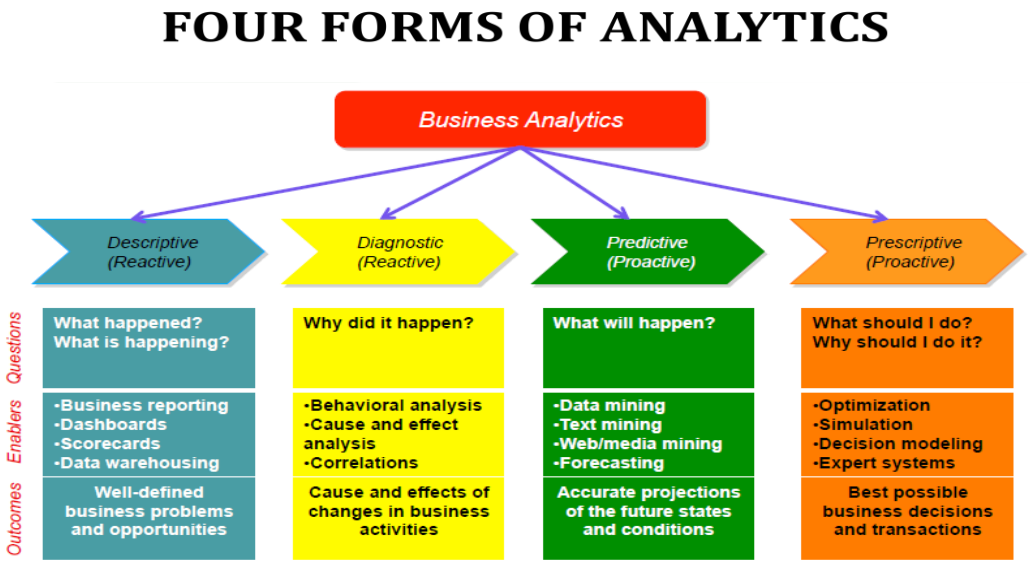

FIGURE 6: ANALYTICS OF BIG DATA

Data analysis always plays a prominent role in the market; the stored data is drilled to draw the unseen facts of it. Machine Learning algorithms play a major role in the process of prediction[11]. There are broadly two categories of machine learning often called Supervised and unsupervised learning, where supervised learning makes out analysis on the gathered sets of data called experience, where we try to sort out the possible occurrences of the current problem looking into the past history of how to handle such similar issues. These can be done using algorithms that support Regression and Classification. The unsupervised category unlike prior one doesn't have any history and we need to derive a structure from the given data where Clustering is used. For Big data it is often preferable to do Feature selection, Classification and clustering.

A. **Feature Selection:** It is referred as a pre-processing step before data mining where we select a subset of features and remove the irrelevant features in order to reduce the computational complexity of the algorithm following approaches like Feature search done using any of the methods like complete, selection, Sequential and random search approaches. Feature evaluation describes a class of labels to describe the relevancy and correlation of one class of variables with another.[1][5] The feature evaluation uses few approaches like Wrapper, filter and hybrid approaches.

B. **Classification:** It is a supervised machine learning approach. There are several traditional classification mechanisms used to determine the data sets that fall under a particular class. A few approaches are like Decision tree (DT) as the name itself specifies it is a tree like structure where the data is organised at different levels. During retrieval process this structure requires a comparatively lower computational time, Support Vector Machine(SVM) This technique classifies the data based on the support

vector and hyper plane, which separates data points in higher dimensional space, This is generally used in combination with another technique to deal with the problem of noisy data. We also use techniques like KNN(K-Nearest Neighbour), Bayesian Methods

C. **Clustering:** It is an unsupervised machine learning technique which makes similarity between data points such that data points within same cluster should be of high similarity, data points between different clusters should have least similarity. The clustering mechanisms are broadly divided under two types namely: Partitioned clustering and Hierarchical clustering.

III. Trends and Techniques of Big Data Analytics:

The most popular and significant platform used across various strategies for big data is none other than the Hadoop, its because of its open source distributed data processing and its capability of processing extremely massive amounts of data by following partitioned datasets allocated to numerous servers across where each big task is handled by partitioned programming and integrating at the end. Hadoop is developed by the APACHE platform and it belongs to the class of NOSQL technologies which also include CouchDB, MongoDB that evolved to integrate the data in unique way [10].

Though Hadoop has its advantages there are still issues that are to be addressed, like Hadoop is challenging to install, configure and administer; and also the individuals with Hadoop skills are not easily found. Numerous vendors like AWS, Cloudera, Hortonworks, MapR technologies [16] distribute the open source hadoop platforms. Much proprietary software like BigInsights, BigML sheets are also available. Further platforms like Cassandra, Oozie, HBase, Pig, Hive, MongoDB are used widely for the database component [4]. Among many of the above mentioned techniques are cloud versions and are open source still there are few challenges that require the focus of the big data scientist. As most of the above mentioned tools are open source their development costs may be low and they may be available to the end users free of cost but there are yet some issues that require to be noticed. In an open platform any additional functionality with respect to a specific requirement of a project; requires professionals and subject experts who have very good skills and complete idea of the corresponding domain. So lack of technical support and security issues are the noticed trade-offs that need to be addressed.

TABLE 1: TOOLS & TECHNIQUES FOR BIGDATA:

Platform/Tool	Description
Hadoop Distributed	HDFS is a Hadoop based cluster for storage of huge data by dividing small parts and store them in distrubuted nodes

Map Reduce	The distribution of tasks onto file system can be done with Map Reduce. Whenever information is gathering
Hadoop Distributed	PIG is a high level platform for creating Map Reduce programs used with Hadoop. The language for this platform is Piglatin. PIG can be used to retrieve any kind of data either structured or unstructured. It is executed in Hadoop.
Hive	Hive is a query language it runs on Hadoop architecture. It is similar to SQL.
Jaql	Jaql is used to work on large datasets wherein query language can be used to process the parallel processing data. Low and high level queries can be retrieved.
Zookeeper	Zookeeper is having the huge infrastructure with various services across different clusters. With synchronization process and parallel processing allows the centralized infrastructure data is handled which is helpful in Big Data.
Hbase	Traditional databases are row-oriented database management systems but Hbase is column oriented. It works on the top of HDFS and it is not like an SQL approach
Cassandra	One of the major and most used DBS is Cassandra. It works on the distributed servers where it requires reliable service and no failure. It is also a non SQL based DBS.
Oozie	Streamlining of workflows with co-ordination of different tasks can be done with Oozie. Moreover it is open
Lucene	The people are familiar with the Java development can use Lucene. The Project wide used in text analytics of Big Data
Avro	Avro can be used for version control and it assist in maintaining the configuration management
Mahout	Mahout is very much useful in providing machine learning algorithms in favor of Big data Analytics. It is a Apache project , which is used to develop distributed systems

Several methodologies are developed to handle with the emerging trends of big data and analyze various hidden information to make a meaningful purpose. The methodology of driving any project of any healthcare application with respect to big data analytics can be done following these steps [12]:

1. Understanding the need of big data project and how it reflects the basic V's i.e. our Concept statement.
2. Developing Proposal concentrating on aspects like problem to be addressed, its importance, big data analytics approach required, and background idea.
3. Methodology to be followed at various levels of Data collection, Feature Selection, ETL and data transformation, Platform or tool to be used, Associations, Patterns, Aggregations, etc.

4. Visualization of the results or model developed as part of the analysis
5. Deployment that mainly focuses on Evaluation and validation, testing phases.

IV. Visualization of processed data models:

As more and more businesses are evolving in the day to day world, data visualization has become an increasingly important component in the analytics of big data. The availability of new in-memory technology and high-performance analytics that use data visualization is providing a better way to analyze data more quickly than ever. Visual analytics enables organizations to take raw data and present it in a meaningful way that generates the most value. There are few points of advice to make a effective visualization they are 1. Do not forget representing Meta Data, 2. Always make the visualization interactive, indulging user involvement to make any further changes.

There are several benefits of data visualization tools that van be viewed in the table below.

TABLE 2: BENEFITS OF DATA VISUALIZATION:

Benefits	Percentages(%)
Improved decision-making	77
Better ad-hoc data analysis	43
Improved collaboration	41
Provide self-service	36
Increased return on	34
Time savings	20
Reduced burden on IT	15

Visualization approaches are used to create tables, diagrams, images, and other intuitive display ways to represent data. Big Data visualization is not as easy as that of the traditional small data sets. The extension of traditional visualization approaches have already been emerged but far from enough. In large-scale data visualization, many researchers use feature extraction and geometric modelling to greatly reduce data size before actual data rendering. Choosing proper data representation is also very important when visualizing big data[20].

A lot of big data visualization tools run on the Hadoop platform. The common modules in Hadoop are: Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce. They analyze big data efficiently, but lack adequate visualization. Some software with the functions of visualization and interaction for visualizing data has been developed like Pentaho: It supports the spectrum of BI functions such as analys1is, dashboard, enterprise-class reporting, and data mining. Flare: An ActionScript library for creating data visualization that runs in Adobe Flash Player. JasperReports: It has a novel software layer for generating reports from the big data storages. Dygraphs: It is quick and elastic open source

JavaScript charting collection that helps discover and understand opaque data sets. Datameer Analytics Solution and Cloudera: Datameer and Cloudera have partnered to make it easier and faster to put Hadoop into production and help users to leverage the power of Hadoop. Platfora: Platfora converts raw big data in Hadoop into interactive data processing engine. It has modular functionality of in-memory data engine. ManyEyes: It is a visualization tool launched by IBM. Many Eyes is a public website where users can upload data and create interactive visualization. Tableau: It is a business intelligence (BI) software tool that supports interactive and visual analysis of data. It has an in-memory data engine to accelerate visualization. Tableau has three main products to process large-scale datasets, including Tableau Desktop, Tableau Sever, and Tableau Public. Tableau also embed Hadoop infrastructure. It uses Hive to structure queries and cache information for in-memory analytics. Caching helps reduce the latency of a Hadoop cluster. Therefore, it can provide an interactive mechanism between users and Big Data applications.

Conclusions and Future Enhancements:

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. Predictive analytics is a key technique that helps in tracing the upcoming scenarios with respect to current symptoms of any health issue. Big data analytics application in healthcare domain helps in diagnosis, prognosis, early intervention and optimal management of the data. There are few trade-offs that are to be handled as part of the future enhancements by the upcoming big data scientist like the technical support with respect to several platforms and patterns of Hadoop architecture, developing Skills and subject Professionals to handle several real time issues[5]. To provide an effective and secured transmission privacy is a major area to be focused[14][15].The visualization strategy also contains major issues like enhancing perceptual and interactive visualization capability. Also concentrate on meeting the need for speed, understanding data, Addressing data quality and displaying meaningful results.

References:

[1] Big data Analytics in Healthcare: A Survey Approach , Dharavath Ramesh1, Member, IEEE , Pranshu Suraj2, and Lokendra Saini3 Department of Computer Science and Engineering Indian School of Mines, Dhanbad,Jharkhand-26004,India,Email:ramesh.d.in@ieee.org1,{pranshusuraj2,lokendras903}@gmail.com

[2] Big Data in healthcare:Challenges and Opportunities, by *Hiba Asri OSER research team, FSTG Cadi Ayyad University Marrakesh, Morocco hiba.asri@gmail.com, Hajar Mousannif LISI Laboratory, FSSM Cadi Ayyad University Marrakesh, Morocco mousannif@uca.ma, 978-1-4673-8149-9/15/$31.00 ©2015 IEEE

[3] Predictive Big Data Analytics in Healthcare, A.Rishika Reddy Computer Science and Engineering Kakatiya Institute of Technology & Science Warangal, India rishika51896@hotmail.com, P. Suresh Kumar Computer Science and Engineering Kakatiya Institute of Technology and Science Warangal, India peddojusuresh@gmail.com, 2016 Second International Conference on Computational Intelligence & Communication Technology, 978-1-5090-0210-8/16 $31.00 © 2016 IEEE DOI 10.1109/CICT.2016.129

[4] Big Data: issues, tools and challenges by Avita Katal, Mohd Wazid, R H Goudar at IEEE 2013. 978-1-4799-0192-0/13/$31.00 ©2013 IEEE

[5] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", IEEE, 46th Hawaii International Conference on System Sciences, 2013.
[6] Big data analytics in healthcare: promise and potential Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3 http://www.hissjournal.com/content/2/1/3
[7] Big Data for Health Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang, Fellow, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4, JULY 2015
[8] Big data analytics ,FOURTH QUaRTeR 2011,By Philip Russom.
[9] Design Principles for Effective Knowledge Discovery from Big Data, Edmon Begoli, James Horey, 2012 Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture
[10] Trends in bigdata analytics Karthik Kambatlaa, ∗,Giorgos Kolli asb ,Vipin Kumar c, Ananth Grama a, E-mailaddresses:kkambatl@cs.purdue.edu(K.Kambatla),gkollias@us.ibm.com (G.Kollias),kumar@cs.umn.edu(V.Kumar),ayg@cs.purdue.edu(A.Grama),J.Parallel Distrib.Computing journal homepage: www.elsevier.com/locate/jpdc.
[11] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, 2012.
[12] M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. W. Treister, Transforming Health Care Through Big Data, Institute for Health Technology Transformation, Washington DC, USA, 2013.
[13] Dembosky A: "Data Prescription for Better Healthcare." Financial Times, December 12, 2012, p. 19; 2012. Available from: http://www.ft.com/intl/cms/ s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK.
[14] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/ wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.
[15] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, "Big Data Privacy Issues in Public Social Media", IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST), 18-20 June 2012.
[16] Improving Map Reduce Performance in Heterogeneous Environments, USENIX Association, SanDiego, CA,2008, 12/2008.
[17] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel Abadi, Alexander Rasin, Avi Silberschatz, HadoopDB: an architectural hybrid of MapReduce and DBMS technologiesforanalyticalworkloads,in:VLDB,2009.
[18] G. N. Forrest, T. C. Van Schooneveld, R. Kullar, L. T. Schulz, P. Duong, andM.Postelnick,"Use of electronic health records and clinical decision support systems for anti microbial stewardship,"Clin.InfectiousDis.,vol. 59, pp. 122–133, 2014.
[19] Big data computing and clouds: Trends and future directions by Marcos D. Assunc¸˜ao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya, Big data computing and clouds: Trends and future directions, J. Parallel Distrib. Comput. (2014), http://dx.doi.org/10.1016/j.jpdc.2014.08.003
[20] Big Data and Visualization: Methods, Challenges and Technology Progress Lidong Wang1,*, Guanghui Wang2, Cheryl Ann Alexander3, Digital Technologies, 2015, Vol. 1, No. 1, 33-38 Available online at http://pubs.sciepub.com/dt/1/1/7 © Science and Education Publishing DOI:10.12691/dt-1-1-7

AUTHOR:

Dr M V BRAMHANANDA REDDY

B.TECH,M.TECH,Ph D(CSE)

DEPT OF COMPUTER SCIENCE AND ENGINEERING

SAI RAJESWARI INSTITUT OF TECHNOLOGY, LINGAPURAM ,PRODDATUR.

Email Id: bramhareddysrit2019@gmail.com

CELL NO: 9000840692