Knowledge Discovery in Medical Big Data by Leveraging the Benefits of Data Mining and Hadoop Approaches

Mrs. Renita Tellis[1],
[1] Assistant Professor,
Computer Science & Engineering,
Mangalore Institute of Technology and Engineering,
Moodbidri, Karnataka, India

Mrs. Lathika J Shetty[2]
[2]Assistant Professor,
Computer Science & Engineering,
Mangalore Institute of Technology and Engineering,
Moodbidri, Karnataka, India

***Abstract:-* In this paper we analyze and reveal the potentials of Data mining and Hadoop technology in retrieving essential medical data which may otherwise be hidden or go dissipated. In today's world, due to increase in population and information in various sectors, data is moving from small data to big data. Healthcare industry generates large amount of data driven by record keeping, compliance and regulatory requirements, and patient care. In developing countries like India with a huge population, providing the benefits of healthcare to a common man at a reasonable price has become a great issue. Accurate knowledge discovery in medical big data leads to confident decision making which can guarantee greater operational efficiency, reduced risk and reduction in cost. This paper gives the involvement of Big Data analytics to render the services of healthcare to everyone at an optimal cost.**

Key Words: Health care in India, HDFS, HIPI, KDD, Medical Big Data, MapReduce, WEKA

1. INTRODUCTION

In this fast paced competitive world, human health is considered to be priceless. Health once lost is difficult to be recovered. Therefore a sincere attempt is made to effectively incorporate the benefits of information technology for healthcare to make the wellbeing of humans a priority.

Healthcare industry consists of humungous amount of data. A methodical procedure for analyzing, storing, processing and validating this data is necessary. Therefore to achieve this goal, major techniques like data mining and hadoop have contributed various forms to deliver applications in the area of healthcare. WEKA is a collection of machine learning algorithms that can be used for data mining tasks in healthcare. However, analyzing healthcare data using hadoop is given more focus in this study since it reduces the cost of services to a common man in the country.

2. MEDICAL BIG DATA

Datasets that distend limits of traditional data processing and storage systems is referred to as Big Data. Big Data includes datasets with sizes which are beyond the ability of commonly used software tools to manage and process data.

Big Data can be described by the following characteristics as illustrated in Fig-1:

- Volume
- Velocity
- Variety
- Veracity

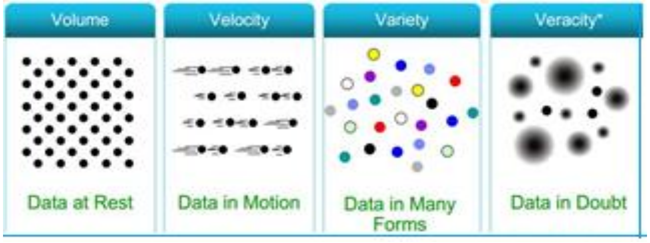

Fig-1: Characteristics of Big Data

Each time a person comes in contact with a healthcare professional, some type of data is produced. This data can be on paper, electronic, or both. Every patient, from birth to death accumulates a wide spectrum of variables of healthcare data as shown in Fig-2.

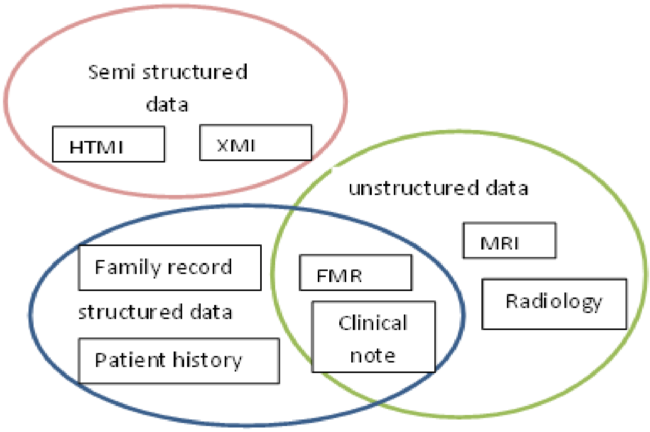

Fig-2: Medical data types collected

3. BIG DATA ANALYTICS IN HEALTHCARE

Healthcare industry generates big amount of data which includes record on symptoms, clinical findings, medications, family history and patient history. Since there is a need to manage this Big Data and to extract potentially required values and hidden knowledge from it, a concept called Big Data Analytics was introduced as illustrated in Fig-3.

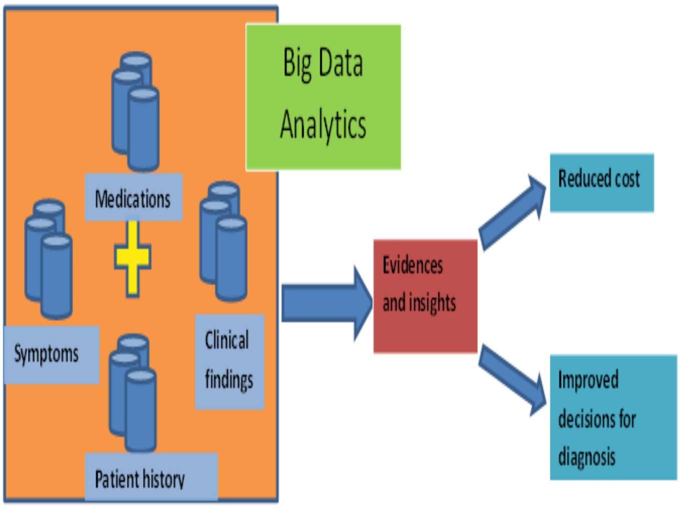

Fig-3: Big Data Analytics in Healthcare

4. KNOWLEDGE DISCOVERY IN MEDICAL BIG DATA

We are drowning in a sea of data but are starving for knowledge. It is necessary to extract useful information from large collection of data stored in an organization database. KDD (Knowledge Discovery in Databases) is a non-trivial process of identifying valid, novel, potentially useful and understandable patterns in data.

5. HEALTHCARE IN INDIA

India takes the second place in world population with the count of 1.25 billion. OECD(Organization for Economic Co-operation and Development)illustrates that India ranks well below the OECD average in terms of health expenditure per capita with spending of only USD 157 compared with an OECD average of USD 3484.Total health spending accounted for only 4.0% of GDP in India, less than half the OECD average of 9.3%.

Table -1: Countries expenditure on healthcare

Country	Total % of GDP spent on Healthcare	Per capita spent on Healthcare (USD)
USA	16.9	8745
AUSTRALIA	9.1	3997
JAPAN	10.3	3649
UK	9.3	3289
KOREA	7.6	2291
INDIA	4.0	157

Table-1 shows the countries expenditure on healthcare. Chart-1 and Chart-2 show the graphical representation of the total % of GDP spent on Healthcare and Per capita spent on Healthcare (USD) respectively.

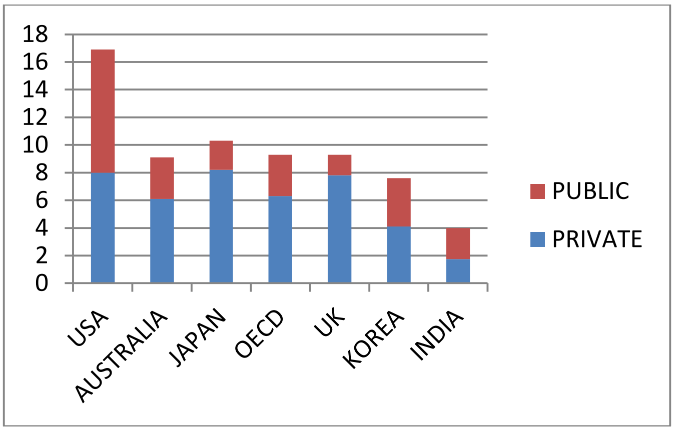

Chart-1: Total % of GDP spent on Healthcare

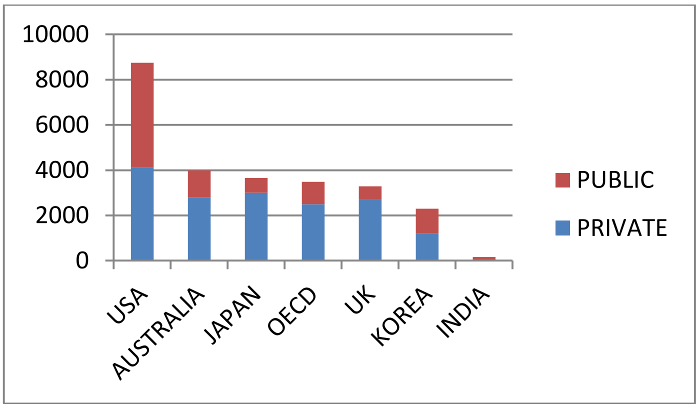

Chart-2: Per capita spent on Healthcare (USD)

From the above statistics one can find that India's population pays more for medical services due to irregularity in storing and updating of healthcare data. Spending out of their pockets for healthcare pushes majority of people into poverty thus hindering the development of India. Therefore it is essential for India to acquire the benefits of technology by incorporating the techniques of data mining and hadoop to gain availability of necessary healthcare data on the desktop and accessibility of useful information on hospitals, medicines and relevant patient information, thus improving the entire system of healthcare in India.

6. TECHNIQUES AND TECHNOLOGY

Here we discuss about the general methodology and the different methods applied on preprocessed healthcare dataset.

6.1 General methodology

The methodology is an iterative sequential process of data collection which is the process of gathering information on the domain of interest. Domains can be business transactions, scientific data, medical and personal data, surveillance video and picture, satellite sensing, World

Wide Web repositories etc. It then continues with data selection, where the target data set of the desired domain is taken into consideration.

The data selected must in one of the four formats given below:

- ARFF(Attribute Relation File Format) has two sections –
    - The header information defines attribute name , type and relations
    - The data section lists the data records
- CSV (Comma Separated Values)
- C4.5 :It requires two separate files
    - Name file: Defines the names of the attributes
    - Data file: Lists the records (samples)
- Binary

This formatted data is given to the next step which includes data preprocessing which is followed by application of analytic method on the preprocessed data. Analytic method can be either data mining or hadoop. The output pattern is then evaluated for knowledge discovery .The final output is displayed in an interface viewable by the common man. The
Fig-4 illustrates the methodology.

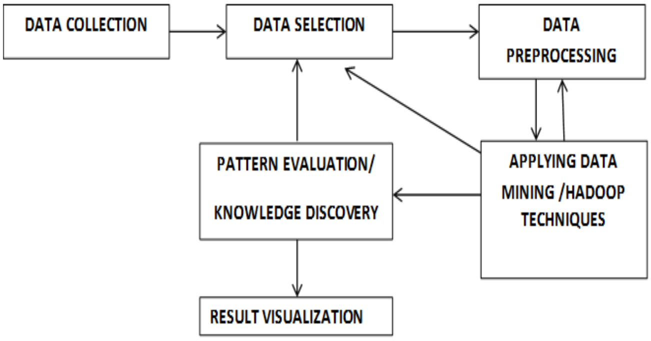
DATA COLLECTION
DATA SELECTION
DATA PREPROCESSING
PATTERN EVALUATION/ KNOWLEDGE DISCOVERY
APPLYING DATA MINING /HADOOP TECHNIQUES
RESULT VISUALIZATION

Fig-4: General Methodology

6.2 Methods applied on preprocessed healthcare data

Here we discuss about the approaches to knowledge discovery in medical Big Data.

6.2.1 Data Mining Techniques

Data mining is a computer based process of analyzing large amount of data and extracting relevant and useful information. It is an important step of KDD process.

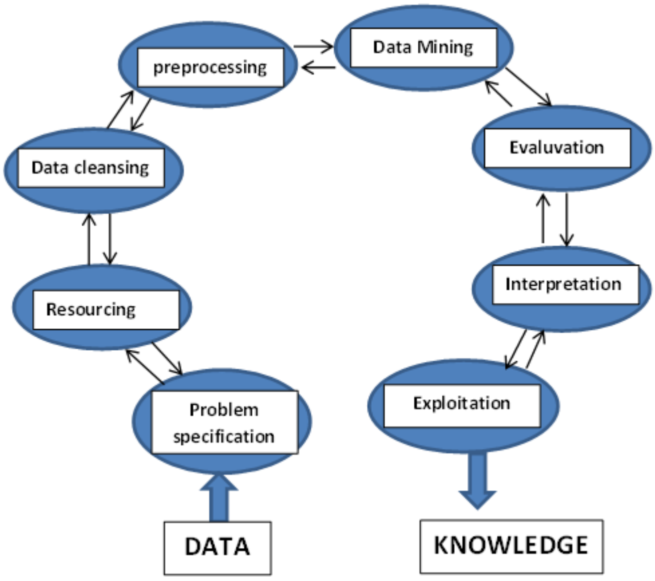

Fig-5: KDD process

Fig-5 shows the following iterative sequential steps

1. Data collection: learning the application domain. Example: Medical and personal data
2. Data selection: creating a target data set which will be subjected to analysis. Example: heart disease dataset from the UCI repository
3. Data pre-processing: The chosen health care datasets are pre-processed to handle problems like noise, missing and inconsistent data. This step transforms data into a form that is presentable to the data mining techniques.
4. Data mining: This involves the task of analyzing the dataset and extracting the data patterns using various data mining algorithms like classification, regression, association and clustering.
5. Pattern evaluation and knowledge discovery: A systematic determination of strictly interesting patterns representing knowledge, is done using criteria governed by a set of standards.
6. Result visualization: It is a final Phase where the knowledge discovered is represented visually to the user to understand and interpret the results.

WEKA (Waikato Environment for Knowledge Analysis) developed at University of Waikato in New Zealand, is an open source collection of data mining algorithms software written in JAVA which is a very useful and efficient application for analysis of Big Data using data mining techniques.
Features comprises of:

- Graphical User Interfaces(including data visualization)
- Comprehensive set of data pre-processing tools, data mining algorithms and evaluation methods
- Environment for comparing algorithms.

6.2.2 Hadoop framework

Hadoop is an open source framework which uses simple programming models to allow storing and processing of Big Data in a distributed computing environment across clusters of computers. It provides a design to scale up from single servers to thousands of machines, each offering local computation and storage. One of the most efficient solution for Big Data Analytics is hadoop.

Hadoop framework consists of the following components:

- HDFS (Hadoop Distributed File System) - It is a block structured distributed file system which holds large amount of Big Data. It provides high throughput access to application data. HDFS uses master/slave architecture. Master consists of a single name node that manages the file system metadata and one or more slave data nodes that store the actual data.
- Hadoop MapReduce- Hadoop runs applications using map reduce algorithm which is a programming framework for distributed computing. Here divide and conquer method is used to break large complex data into smaller units and process them as shown in Fig-6.

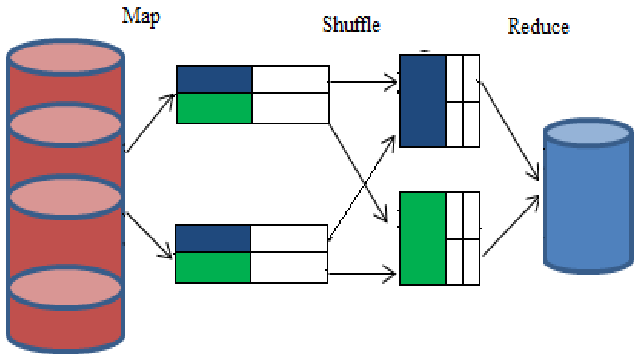

Fig-6: MapReduce

Fig-7 illustrates that MapReduce framework consists of a single master Job tracker and one slave Task tracker per cluster node. The slave Task tracker executes the task as directed by the master and provides task status information to the master periodically.

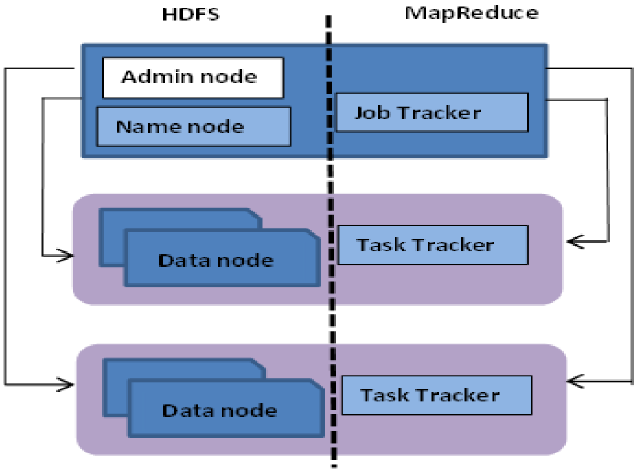

Fig-7: Hadoop Components

- Other hadoop components :
  - HBase: HBase is open source, Non-relational, distributed database system written in Java. It runs on the top of HDFS. It can serve as the input and output for the MapReduce.
  - Pig: Pig is high-level platform where the MapReduce programs are created which is used with Hadoop. It is a high level data processing system where the data sets are analyzed that occurs in high level language.
  - Hive: Hive is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis.
  - Sqoop: Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.
  - Avro: Avro is a data serialization system and data exchange service. It is basically used in Apache Hadoop. These services can be used together as well as independently.
  - Oozie: Oozie is a java based web-application that runs in a java servlet. Oozie uses the database to store definition of Workflow that is a collection of actions. It manages the Hadoop jobs.
  - Chukwa: Chukwa is a data collection and analysis framework which is used to process and analyze the large amount logs. It is built on the top of the HDFS and MapReduce framework.
  - Flume: Flume is high level architecture which focused on streaming of data from multiple sources.
  - Zookeeper: Zookeeper is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc.

HDP (Hortonworks Data Platform) is the most efficient application for healthcare involving hadoop and Big Data. This platform is designed to deal with data from many sources and file formats. Fig-8 shows the architecture of HDP

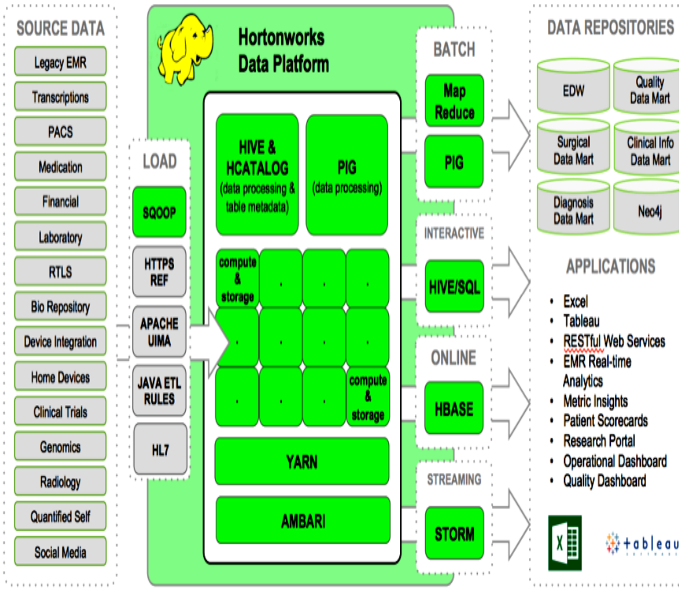

Fig-8: Hortonworks Data Platform

HIPI (Hadoop Image Processing Interface)
Processing medical images is one of the major challenges faced in healthcare industry. Hadoop provide solutions to analyze medical images in order to gather potentially useful data to give right diagnosis. HIPI provide an API for image processing in distributed computing environment.Fig-9 illustrates the functioning of HIPI.

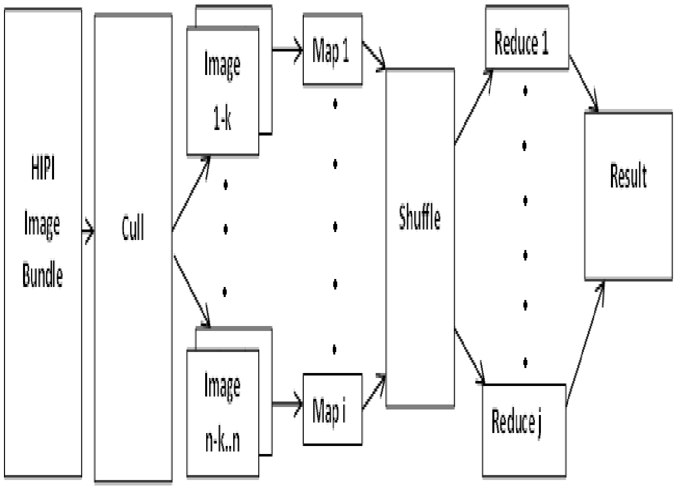

Fig-9: HIPI

The input to a HIPI program is a HipiImageBundle (HIB). HIB is a single file on the HDFS, the file being a collection of images. This is followed by the culling step which is filtering of images. The output of the culling step is assigned to individual map task. The data produced by the map stage is given to reduce task according to map reduce shuffle algorithm. Finally these reduced tasks are executed in parallel and their output is collected and returned to HDFS.

7. FUTURE RESEARCH DIRECTIONS

Healthcare being a critical issue requires favorable amount of analysis to produce precise data in order to provide right diagnosis. Even though our analysis on data mining and hadoop on healthcare helps in making accurate decisions on diagnosis, reducing time, cost and increase in analytic flexibility ,the following obstacles fall in the pathway of yielding optimum results.

- Real-time capturing of data during emergency.
- Conveying Real-time data to the right healthcare professional for accurate analysis.
- Lack of technical knowledge of healthcare professionals making them uncomfortable to use all the facilities provided by information technology.
- Lack of proper basic infrastructure.
- Difficulty in creating IT infrastructure needed to pursue Big Data analytics and related projects.
- Strict regulatory requirements that upholds patient privacy and data security.
- Lack of standardized healthcare data with incompatible formats.

8. CONCLUSION

Big Data Analytics has the potential to transform the way health care professionals use information technology to gain insight from medical Big Data and make right decisions. The challenges highlighted above must be addressed in order to contribute to the healthcare industry. The techniques of data mining and hadoop can allow easy accessibility and availability of healthcare at lower cost. Thus helping the human population reap the benefits of Big Data analytics and taking the country towards development.

REFERENCES

[1] Deepak Kumar, B. 2011. “Evaluation of the Medical Records System in an Upcoming Teaching Hospital—A Project for Improvisation”, Journal of Medical Systems.

[2] http://www.oecd.org/els/health-systems/Health-at-a-Glance-2013.pdf

[3] http://www.oecd.org/els/health-systems/Briefing-Note-INDIA-2014.pdf

[4] Frawley, W., Batheus, C., 1991. Knowledge Discovery in Databases: An Overview. In Piatetsky-Shapiro, G. and Frawley, W. (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, MA, ppl-27.

[5] Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Fco., CA., USA.

[6] Witten, I. H., Frank, E. 2000. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA USA, 371 pp.

[7] Zhou, Z. H., 2003. Three perspectives of data mining. Artificial Intelligence, N 143(1), pp.139-146.

[8] http://searchcloudcomputing.techtarget.com/definition/ Hadoop

[9] Siegel, J. and Perdue, J. Cloud Services Measures for Global Use: The Service Measurement Index (SMI)," SRII Global Conference (SRII), 2012 Annual, vol., no., pp.411, 415, 24-27 July 2012

[10] Nong Y., The Handbook of Data Mining. Lawrence Earlbaum Associates, 2003.

[11] Weka, “Data Mining Machine Learning Software, [Online] Available: http://www.cs.waikato.ac.nz/ml/weka
[12] Sagiroglu, S.; Sinanc, D., (20-24 May 2013),”Big Data: A Review”
[13] https://dzone.com/big-data
[14] https://en.wikipedia.org/wiki/Hortonworks
[15] https://cs.ucsb.edu/~cmsweeney/papers
[16] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W.,(18-22 Dec.,2012) , “Shared disk big data analytics with Apache Hadoop”