Internship Training

At

Aguai Solutions,

Bangalore

On

Understanding the Concept of Big Data Analytics In Healthcare

By

Dr. Mamta Gupta

PGDHM

2012-14

International Institute of Health Management Research

New Delhi

Internship Training

At

Aguai Solutions,

Bangalore

On

Understanding the Concept of Big Data Analytics In Healthcare

By

Dr. Mamta Gupta

Under the guidance of

Dr. Anandhi Ramachandra

Post Graduate Diploma in Hospital and Health Management

2012-2014

International Institute of Health Management Research

New Delhi

The certificate is awarded to

DR. MAMTA GUPTA

In recognition of having successfully completed her

Internship in the department of

HEALTHCARE IT

And has successfully completed her Project on

UNDERSTANDING THE CONCEPT OF BIG DATA ANALYTICS IN HEALTHCARE

At

AGUAI SOLUTIONS, BANGALORE

Duration – February 14TH to May 13TH, 2014

She comes across as a committed, sincere & diligent person who has a Strong drive & zeal for learning

We wish her all the best for future endeavours!

Training & Development

Zonal Head-Human Resources

TO WHOMSOEVER MAY CONCERN

This is to certify that Dr. Mamta Gupta is a student of Post Graduate Diploma in Hospital and Health Management (PGDHM) from International Institute of Health Management Research, New Delhi has done dissertation at Aguai Solutions, Bangalore from February 14, 2014 to May13, 2014.

The Candidate has successfully carried out the study designated to him during internship training and his approach to the study has been sincere, scientific and analytical. The Internship is in fulfilment of the course requirements. I wish her all success in all his future endeavours.

20/5/2014

Dr. A.K. Agarwal
Dean, Academics and Student Affairs
IIHMR, New Delhi

Dr. Anandhi Ramachandran
Faculty
IIHMR, New Delhi

Certificate of Approval

The following dissertation titled "UNDERSTANDING THE CONCEPT OF BIG DATA ANALYTICS IN HEALTHCARE" at "AGUAI SOLUTIONS, BANGALORE" is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of **Post Graduate Diploma in Health and Hospital Management** for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the dissertation only for the purpose it is submitted.

Dissertation Examination Committee for evaluation of dissertation.

Name	Signature
Dr. Anandh Ramadoss	
DR. ABHIJIT CHAKRABORTY	07/05/14

Certificate from Dissertation Advisory Committee

This is to certify that **Dr. Mamta Gupta,** a graduate student of the **Post- Graduate Diploma in Health and Hospital Management** has worked under our guidance and supervision. He/ She is submitting this dissertation titled UNDERSTANDING THE CONCEPT OF BIG DATA ANALYTICS IN HEALTHCARE at AGUAI SOLUTIONS, BANGALORE in partial fulfilment of the requirements for the award of the **Post- Graduate Diploma in Health and Hospital Management.** This dissertation has the requisite standard and to the best of our knowledge no part of it has been reproduced from any other dissertation, monograph, report or book.

Dr .Anandhi Ramachandran
Institute Mentor Name,
Assistant Professor, IIHMR
Designation, Organization

Mr. Bimlesh Gundaro
Organization Mentor Name
CEO, Aguai Solutions
Designation, Organization

INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH, NEW DELHI

CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled UNDERSTANDING THE CONCEPT OF BIG DATA ANALYTICS IN HEALTHCARE in AGUAI SOLUTIONS, BANGALORE submitted by DR. MAMTA GUPTA, Enrolment No. PG/12/044 under the supervision of DR.ANANDHI RAMACHANDRAN for award of Postgraduate Diploma in Hospital and Health Management of the Institute carried out during the period from 14th February to 13th May, 2014 embodies my original work and has not formed the basis for the award of any degree, diploma associate ship, fellowship, titles in this or any other Institute or other similar institution of higher learning.

Mamta

Signature

FEEDBACK FORM

Name of the Student: DR. MAMTA GUPTA

Dissertation Organisation: AGUAI SOLUTIONS PVT. LTD

Area of Dissertation: UNDERSTANDING THE CONCEPT OF BIG DATA ANALYTICS IN HEALTHCARE

Attendance: REGULAR

Objectives achieved:
1. SECONDARY RESEARCH
2. KNOWLEDGE ENHANCEMENT
3. BIG DATA TOOLS AWARENESS

Deliverables: PROJECT REPORT

Strengths:
1. COMMUNICATION
2. COMMITMENT
3. FAST LEARNER

Suggestions for Improvement:

Signature of the Officer-in-Charge/ Organisation Mentor (Dissertation)

Date: 15/05/2014
Place: Bangalore

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete my internship. A special gratitude I give to, **Mr. Bimlesh Gundaro** (CEO, Aguai Solutions) whose contribution in stimulating suggestions and encouragement, helped me to coordinate my training especially in writing this report.

I would like to acknowledge my gratitude to **Dr. Niaz Ahmed** (Sr. Healthcare Consultant, Aguai Solutions) for his guidance during internship period.

Furthermore I would also like to acknowledge with much appreciation the crucial role of **Dr.Anandhi Ramachandran**, faculty IIHMR and my mentor who gave the support necessary to complete the task. Special thanks to **Dr. Abhijit Chakrabarty**, faculty IIHMR whose teaching was of great help during the period. I have to appreciate the guidance given by my mentor as well as panel especially in my project presentation that has helped me to complete this report.

ABBREVIATIONS

ACO- Accountable care organization

BD-Big Data

BI- Business Intelligence

CME- Continue medical education

CRM- Customer relationship management

DBMS – Database management system

DW- Data warehouse

EHR- Electronic Health Record

EMR- Electronic Medical Record

ETL- Extract-transform-load

ETRI- Electronics and Telecommunications Research Institute

HDFS- Hadoop distribution file system

HEDIS-Healthcare effectiveness data & information set

HIE- Health information exchange

HIPAA- Healthcare insurance portability and accountability act

HIS-hospital information system

HIV- Human immune-deficiency virus

HL7- Health level 7

HPI- History of present illness

ICD- International classification of disease

IT- Information Technology

M- Mean

OS- Operating system

PDE- Person event data environment

PH- Public health

PQRS- Physician quality reporting system

R&D- Research & Development

SD- Standard deviation

VA- Visual analytics

WSJ-Wall street journal

LIST OF TABLES

LIST OF FIGURES

INTERNSHIP REPORT

ORGANIZATION PROFILE

Aguai Solutions is a Company that Delivers Technology & Talent Management Solutions for Startups and growth oriented companies to reduce complexity and help achieve excellence. Aguai Solutions specializes in leveraging its vast Leadership experience to deliver value to its esteemed global customers. Its unique Predictive Delivery Model built upon Predictability, Transparency & Collaboration helps to "Deliver Business Outcomes" consistently. Focus is on delivering services through “leadership”.

Aguai Solutions has helped organizations, ranging from disruptive High Technology startups, SMBs to Healthcare IT companies, leverage knowledge capital, consulting and branding expertise to scale their business, streamline delivery, and delight customers, the world over.

VISION

To become the most preferred business partner to our customers through leadership in our actions, values and social responsibility.

MISSION

To be a world class organization in enabling clients to become Leaders in their industry

NDUSTRIES

It is involved in Retail, Healthcare and Information technology industries.

RETAIL INDUSTRY

Ecommerce firms are setting newer standards in retail, offering compelling price, assortment and service, thus poaching business from traditional in-person shopping. Services it provides are;

1. **BUSINESS**: It helps companies to maximize ROI on these partnerships and manage collaboration successfully. It provide specific services;

- Strategic Planning Process
- Business Strategy Definition & Review
- Marketing Strategy Definition & Review
- Operational Strategy Definition & Review
- Digital Marketing Strategy & Execution

2. **TALENT DEVELOPMENT**: Aguai Solutions define talent solutions and engage its people to achieve business objectives. It provides specific services;

- Talent Acquisition Strategy
- Talent Development road map & Execution

3. **TECHNOLOGY**: Aguai Solutions provide cutting edge technology vendors and deliver best –in-class technological and support skills. It deliver specific services;

- Technology Vendor evaluation & recommendation
- Technology recommendation
- Program Management
- Technology vendor management
- Project management

HEALTHCARE IT INDUSTRY

Its focus is on emerging economies, personalized medicine & technological advances, a significant ageing population, rising costs, global pandemics, environmental changes, evidence-based medicine, shortage of medical staff, payers' influence over treatment decisions, philanthropy, vaccination, and medical tourism. Services it provides are;

1. BUSINESS: Aguai offers consulting services in the areas of Business Strategy and Management, Clinical and Operational Process Optimization, Technology and Infrastructure Planning and Integration. Specific Services:

PAYERS

- ICD-10 Transition
- Healthcare Reform Compliance
- Health Analytics & Business Intelligence
- Claims Processing

HEALTHCARE PROVIDERS

- Regulatory Compliance
- ICD-10 Transition
- Healthcare Reform

- Self Service & Patient Engagement
- Hospital Performance Management
- EHR & HIE Implementations
- Clinical Analytics & Business Intelligence
- Enterprise Applications
- Outsourcing

2. TECHNOLOGY: Aguai Solutions has developed and implemented transformational IT solutions focused at Healthcare. From Electronic Health records to Hospital Management Systems, constantly innovated to set benchmarks for effective and affordable care with a strong IT foundation. Specific Services:

- Solution Consulting/ Recommendations
- Product/Solution Road map Definition/Review
- Patient Care Solutions (HIS, EMR, mHealth)
- HL7 interface development
- Custom Application Development & Maintenance
    - Migration
    - New Application Development
    - Maintenance
- Product Engineering Transformation (using SCRUM)

3. TALENT DEVELOPMENT: Aguai Solutions enables workforce transformation through leadership training. It helped people develop management solutions that help organizations maintain a focus on providing the highest quality in patient care. Specific Services:

- Strategic Talent Development Strategy Definition & Review
- Healthcare Domain Training
- Leadership Programs
- Patient Care Programs

INTRODUCTION

Data is a raw form of representation of concepts, figures, facts, etc. Data is a pool from where meanings can be derived if used appropriately. It is a source from where meaningful values can be delivered. When this data is organized, it becomes meaningful called information. For better business decisions information is used.

The new data types, being captured by various enterprises called multi-structured data. These include:

- Semi-structured
- Unstructured (document, image, video etc)
- Sensor and machine generated data.

All the type of data gets captured in the databases, the store house of data. In today's world data is captured from everywhere, be it banks, retail market, social networking, media, businesses, etc in the form of audio, video, images, medical 2D, 3D imaging, etc. It is growing beyond terabytes.

INFORMATION

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

KNOWLEDGE

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts

DATA WAREHOUSE

Integrating data from one or more disparate sources creates a central repository of data, a data warehouse (DW). The typical extract-transform-load (ETL)-based data warehouse uses staging, data integration, and access layers to house its key functions.

The data stored in the warehouse is uploaded from the operational systems. The data may pass through an operational data store for additional operations before it is used in the DW for reporting. The integrated data source systems may be considered to be a part of a distributed operational data store layer.

The definition of the data warehouse focuses on data storage. The main source of the data is cleaned, transformed, cataloged and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support (Marakas & O'Brien 2009).

A data warehouse maintains a copy of information from the source transaction systems. This architectural complexity provides the opportunity to:

- Congregate data from multiple sources into a single database so a single query engine can be used to present data.
- Mitigate the problem of database isolation level lock contention in transaction processing systems caused by attempts to run large, long running, analysis queries in transaction processing databases.
- Maintain data history, even if the source transaction systems do not.
- Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger.
- Improve data quality, by providing consistent codes and descriptions, flagging or even fixing bad data.
- Present the organization's information consistently.
- Provide a single common data model for all data of interest regardless of the data's source.
- Restructure the data so that it makes sense to the business users.
- Restructure the data so that it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems.
- Add value to operational business applications, notably customer relationship management (CRM) systems.(wikipedia)

DATA-MART

It is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Sales, Finance, or Marketing. Data marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.

DATABASE

It is an organized collection of data. The data are typically organized to model relevant aspects of reality in a way that supports processes requiring this information.

DATABASE MANAGEMENT SYSTEM (DBMS)

DBMS is a collection of programs that enables you to store, modify, and extract information from a database.

DBMSs are specially designed software applications that interact with the user, other applications, and the database itself to capture and analyze data. A general-purpose DBMS is a software system designed to allow the definition, creation, querying, update, and administration of databases.

DATA MINING

It is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data

mining is the process of finding correlations or patterns among dozens of fields in large relational databases

Data mining is related to

- Provides the link between evolving separate transaction and analytical systems in information technology
- Software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

It sought four types of relationships, which are

- **CLASSES**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **CLUSTERS**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **ASSOCIATIONS**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **SEQUENTIAL PATTERNS**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.

- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

DIFFERENCE BETWEEN DATA MART AND DATA WAREHOUSE

A data warehouse, unlike a data mart, deals with multiple subject areas and is typically implemented and controlled by a central organizational unit such as the corporate Information Technology (IT) group. Often, it is called a central or enterprise data warehouse. Typically, a data warehouse assembles data from multiple source systems. Below is the table showing difference between data warehouse and data mart;

CATEGORY	DATA WAREHOUSE	DATA MART
SCOPE	Corporate	Line of Business (LOB)
SUBJECT	Multiple	Single subject
DATA SOURCES	Many	Few
SIZE (TYPICAL)	100 GB-TB+	< 100 GB
IMPLEMENTATION TIME	Months to years	Months

Table no. 1: Difference between Data warehouse and data mart

DIFFERENCE BETWEEN DATABASE AND DATA WAREHOUSE

The difference between database and data warehouse is given in following table;

CATEGORY	DATABASE	DATA WAREHOUSE
PURPOSE	Data retrieval and management	Data analysis and decision making
SYSTEM/APPLICATION	OLTP(Online Transaction processing system)	Data mining tools, reporting tools
FORMAT	Normalized, relational database	De-normalization, integrated, subject oriented, multi-dimensional arrays
TIME FRAME	Current/ real-time	Historical

Table no. 2: Difference between Database and Data warehouse

DIFFERENCE BETWEEN DATAMINING AND DATA WAREHOUSE

The primary purpose behind both these functions is to provide the tools and methodologies to explore the patterns and meaning in large amount of data. Data warehouse contains a vast number of database tables. These tables are organized to work with each other in a logical, systematic way. The maintenance of these tables is essential to the continuing operation and accuracy of the data warehouse.

CATEGORY	DATA MINING	DATA WAREHOUSE
PURPOSE	Finding patterns, models and forecasting	Storage
FUNCTION	Analyzing data/spot patterns	Extraction and organization of the data in Central repository
RUNS ON	Statistical analysis	Queries and reporting

Table no. 3: Difference between Data mining and Data

DATA USAGE BY HEALTHCARE SETUP FOR BUSINESS SOLUTIONS

As the health sector is going green, making a transition from paper to electronic form the volume of data has also increased. To manage the growing volume and use it for business decision making the data should be processed and analyzed to achieve quality.

Next step was to manage the information derived from data. Information management was done using Business intelligence (BI) tool for optimal use of resources and infrastructure, finding gaps, business solutions, decision making, etc.

All type of information can be derived from BI with the fly. It helps to evaluate performance of hospitals/health centers/clinics. Periodic reporting of services delivered getting generated, using business intelligence tools. Reports are visualized by all access users. BI applications include components like spreadsheets, charts, dashboards, tabular reports, etc.

The whole process of data analysis goes along various steps like data capturing, storage and processing, transferring, sharing, retrieving, analyzing and visualizing/deriving patterns.

The volume of data is growing in health sector as like other sectors. The role of big data analysis came into play and being adopted by various industries. Now it is health sector which has to understand and adopt big data analysis.

BIG DATA

It is a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis –IDC.

It is collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications-Wikipedia.

Big data is distinguished by 4V's:

1. **VOLUME**: The volume of data from private as well as public healthcare sector is increasing tremendously. The volume of data to deal with growing beyond terabytes due to drive for going digitally from paper work in healthcare. A scalable solution like big data needs to be implemented with growth of healthcare data. Analytics will be used to derive relevance out of large growing data to support decision making. Moreover, about 85% of data is unstructured.

2. **VARIETY**: The data comes in various formats/forms, can be structured or unstructured. The data coming from many sources like text, video, audio, sensor data etc. All the data from different sources are used and analyzed in big data analytics.

3. **VELOCITY**: Data is streaming with speed and need to analyze in real-time. A quick response is required to deal with real-time data which is handled by big data analytics.

4. **VERACITY**: The unreliable data or data which can't be trusted is veracity. The cause can be its source itself, entry error, redundancy, etc. Cleaning of noisy data is used on big data through data mining processes to reduce the noisy data.

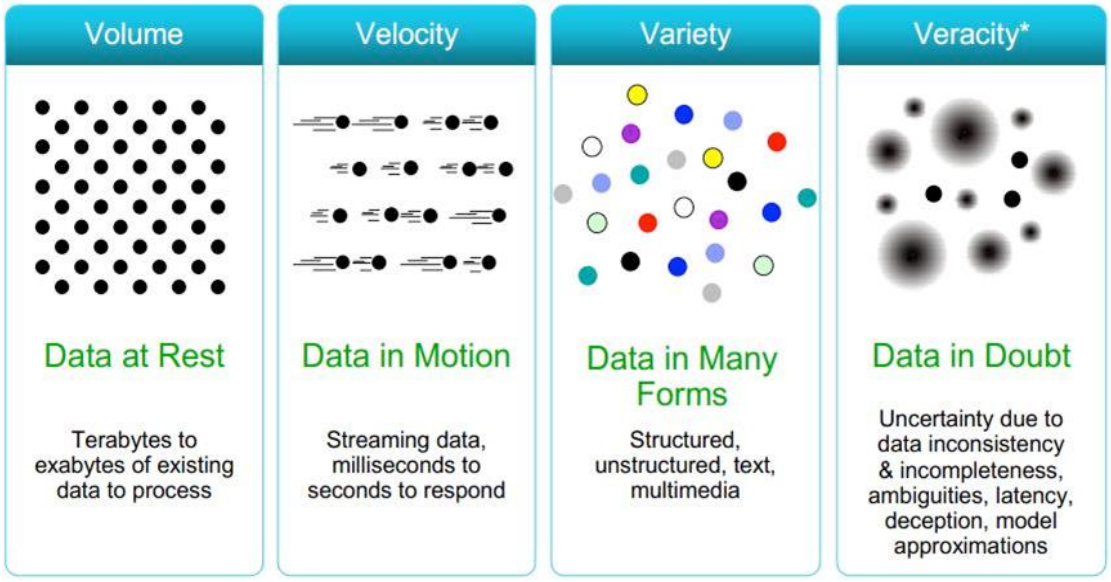


Figure 2-4v's of big data

DATA ANALYSIS

With the e-commerce era, the variety and magnitude of data getting increased. There is a revolution shift to transact data in e-form. Every sector is facing a challenge to manage and utilize the data appropriately.

The structured or unstructured data holds a vast potential for exploring business solutions and opportunities. Enterprises are working really hard to use the data for predicting future solutions. The amount of data getting exploded with variety makes it important to work toward data management.

Now when pay as you go likes cloud concept has entered the e-market, database limitation is not a constraint. Although the multi-dimensional data making it complex to manage appropriately. Hence, data need to be analyzed to find hidden treasure it folds.

Analytics is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Data integration is a precursor to data analysis.

There is different form of analysis;

I. **ARTIFICIAL NEURAL NETWORKS**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

II. **GENETIC ALGORITHMS**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

III. **DECISION TREES**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

IV. **RULE INDUCTION**: The extraction of useful if-then rules from data based on statistical significance.

V. **DATA VISUALIZATION**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

It is important to take the measurement levels of the variables into account for the analyses, as special statistical techniques are available for each level:

- Nominal and ordinal variables
- Frequency counts (numbers and percentages)
- Associations
- circumambulations (cross-tabulations)
- hierarchical log linear analysis (restricted to a maximum of 8 variables)
- log linear analysis (to identify relevant/important variables and possible confounders)
- Exact tests or bootstrapping (in case subgroups are small)
- Computation of new variables
- Continuous variables
- Distribution
- Statistics (M, SD, variance, skewness, kurtosis)
- Stem-and-leaf displays
- Box plots

BIG DATA ANALYTIC PROCESS

The analytic process in big data analytics initiate with collecting un-structured or semi-structured data. Data which is still un-modeled is cleaned to make it free from category of noisy data. After cleaning of unwanted and noisy data including data redundancy data is explored. In the next step of ongoing process, data is modeled or structured in set format. Then the model of whole dataset is used to give insight for further decision making.

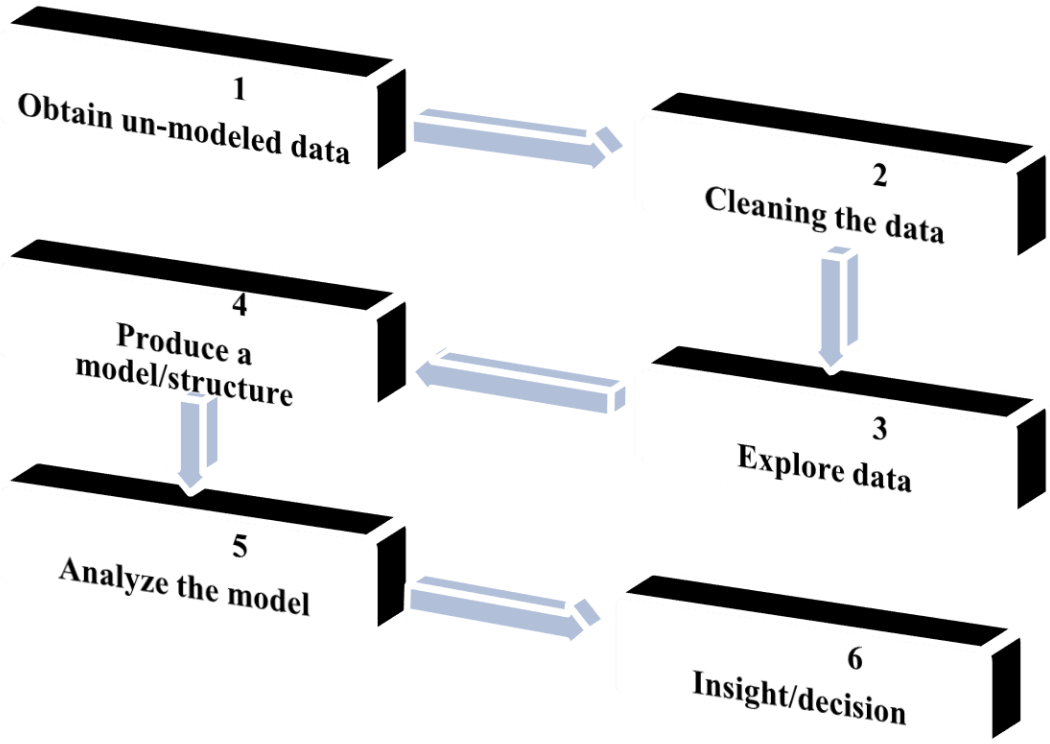


Figure 2- Analytic process

BIG DATA ANALYTICS: AN EMERGING TECHNOLOGY

It is an emerging technology being used for business benefits in healthcare. It uses structured as well as unstructured data and tackles with large volume of data in real-time for analytics. Various organizations plunged into the market to provide solutions for large growing data in healthcare for better decision making. In 2012, The Wall Street Journal had pointed toward big data and explained it as one of the potential areas of application for healthcare.

Figure 3- WSJ

RATIONALE OF STUDY

Healthcare is a healing soul of the society. Emergence of new healthcare IT products aimed at delivering quality care services to the patient. With the adoption of EMR/EHR among healthcare physicians and providers, the dimension of patient's information is increasing manifolds. For instance, a patient record with his name, age, gender, contact no., address, chief complain, HPI, past history, treatment plan, investigations, follow ups, allergies, immunization, minor/major procedures, examination note etc. has to be documented and with each visit the dimensions are increasing tremendously and number of patients also increasing either due to shift of disease patterns or emergence of new diseases or due to other reasons.

This calls for a challenge to tackle the upcoming load of data and at the same time opportunities to explore out of that data: patterns, predictions, decision making for optimal utilization of resources, future benefits, etc via a good analytical tool..

Big data analysis helps to get along with emerging threat toward managing different form of data. It gives solutions required to have better insight for any healthcare setup. Apart from other benefits, it will contribute toward R&D. In R&D, big data analysis help explore human genomic with a pool of available data for further learning.

Big data analysis is still at infancy stage in healthcare. It is a need of an hour for getting better insight/solutions for various healthcare setups.

LITERATURE REVIEW

Big Data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data in a single data set.

In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data. In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Additionally, a new V "Veracity" is added by some organizations to describe it.

The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.999% of these streams, there are 100 collisions of interest per second.

- As a result, only working with less than 0.001% of the sensor stream data, the data flow from all four LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication.
- If all sensor data were to be recorded in LHC, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly 500exabytes per day, before replication. To put the number in perspective, this is equivalent to 500 quintillion (5×1020) bytes per day, almost 200 times higher than all the other sources combined in the world.

  "Big Data" has increased the demand of information management specialists in that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms only specializing in data management and analytics. In 2010, this industry on its own was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

  Potentiality of big data in the medical sector: focus on how to reshape the healthcare system – This study was conducted to explore the use of big data for reducing healthcare concern such as improvement in healthcare systems, appropriate treatment plan, etc. by Creative Future Research Laboratory, ETRI in Korea, 2013. This study explores challenges govt. and healthcare stakeholders facing and opportunities presented by big data. Through this study, it was found that big data applications could help healthcare stakeholders for deploying big data to resolve healthcare issues. In this study, percentages of

persons in families with selected financial burdens related to medical care were also calculated.

The challenge of big data in public health: an opportunity for visual analytics – In this study the efficiency and effective use of public health data was determined, by Western university, Canada in 2014. The extent to which PH stakeholders can address healthcare concerns while dealing with other work activities was determined. Stakeholders interact with data and do cognitive activities like analytical reasoning, decision-making, interpreting, and problem solving. Performing these activities with big data was a challenge as concerns like data's volume, variety, velocity, and veracity were growing for PH stakeholders. The study suggested the use of visual analytics (VA) tools, a category of computational tools that integrate data analytics with interactive visualizations, to facilitate the performance of cognitive activities involving big data. During the study, the potential benefit of incorporating VA tools into PH practice was demonstrated.

The role of big data and advanced analytics in drug discovery, development and commercialization – The study was conducted by McKinsey & Company, 2014. In this study role of big data and advanced analytical methods and technologies used to interpret it. The trend of analytical technology seen in revolutionary fields of biology by McKinsey & Company, Berlin, Germany; medicine by McKinsey & Company, Hamburg, Germany and healthcare by

McKinsey & Company, New York, New York, USA. With new types of data and tools available, analysis of big data is done to see emerging opportunity for smarter and more effective discovery, development, and commercialization of innovative biopharmaceutical drugs.

Big data, advanced analytics and the future of comparative effectiveness research – The study was conducted by Pfizer, Inc., USA in Mar, 2014. Healthcare with the concept of big data, catalyzed by the changing regulatory and competitive environments, fueled by growing adoption of EHR, as well as efforts to integrate medical claims, EHR and other novel data sources. This study assessed that Healthcare organizations have to invest in new hardware, software and skilled individuals for adopting big data and advanced analytics. Analytics will help to revolutionize comparative effectiveness research.

The Person-Event Data Environment: leveraging big data for studies of psychological strengths in soldiers – This study was conducted by Research Facilitation Team, Army Analytics Group Monterey, USA in 2013. United States Army maintains numerous electronic databases with information on more than one million Active-Duty, National Guard soldiers, their family members, and Army civilian employees. The Person-Event Data Environment (PDE) was created to unify disparate Army databases in a secure cloud-based enclave. This study provides an overview of the uses of the PDE to perform command surveillance and policy analysis for Army leadership. Through PDE the role of

psychological assets in major cost drivers such as medical expenditures both during deployment and stateside, drug use, attrition from basic training, and low reenlistment rates were assessed. The study highlights the confluence of both economic and behavioral science perspectives elucidating empirically-based studies examining relations between psychological assets, health, and healthcare utilization.

OBJECTIVES

GENERAL OBJECTIVE:

- To understand the need of big data analytics in healthcare

SPECIFIC OBJECTIVES:

- To understand big data
- To understand need of big data analysis in healthcare
- To understand various opportunistic areas in healthcare
- To know various key players of big data
- To understand various challenges and benefits of big data analysis
- To make people understand big data analysis in healthcare

METHODOLOGY

The purpose of this study is to understand how data is tremendously growing and the need to analyze all data sets for future insight. To understand what analysis can potentially do with existing pool of data especially in healthcare.

RESEARCH METHODOLOGY

STUDY DESIGN: A descriptive methodology was used for this study. Big data analytics in healthcare is still in its infancy stage. All the available data was used to understand its need, finding opportunistic areas, challenges, exploring various available platforms, etc.

During the study, all relevant uses of big data and its analysis in healthcare are identified. Key players in the market are identified.

DATA COLLECTION: The data collection source was secondary. All the available data from secondary source was used to get better understanding of big data and its growing need in healthcare sector. Various available companies' white papers were consulted to get to know big data analysis in healthcare better. Many available documents were also looked into for better insights.

SAMPLE: Online available white papers that of IBM, KPMG, etc were used in this study. Other than white papers, documents, articles, blogs, etc. were used as well in order to have more and more big data information related to healthcare.

RESULT

As the health sector is going green, making a transition from paper to electronic form the volume of data has also increased. To manage the growing volume and use it for business decision making, the data should be processed and analyzed to achieve quality.

Data should be managed properly to derive information and thereafter convert it into knowledge and wisdom for business purposes. Proper processing of available data will be helpful for better insights.

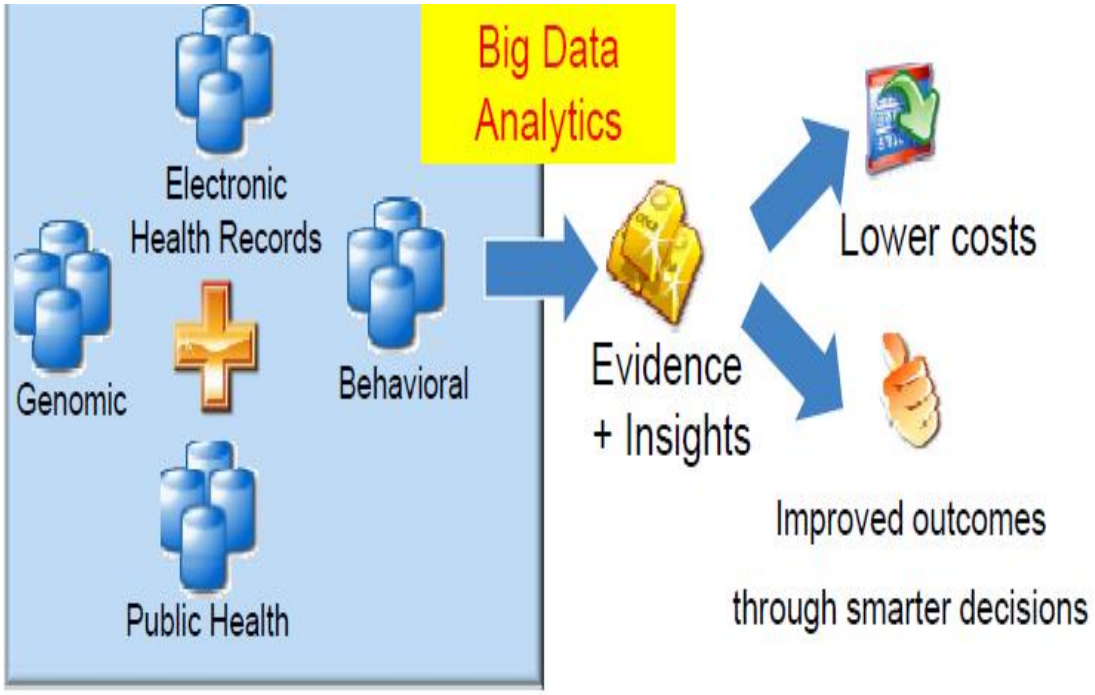
The whole process of data analysis goes along various steps like data capturing, storage and processing, transferring, sharing, retrieving, analyzing and visualizing/deriving patterns.

Information management can be done using Business intelligence tool for optimal use of resources and infrastructure, finding gaps, business solutions, decision making, etc. In business intelligence process of analysis comes up with future beneficiary outputs.

Business intelligence did not analyze whole data set with the volume exceeding beyond terabytes. In this stage of big data, analysis can be done using big data analytical tools. Data can be retrieved from a data warehouse and whole data set can be analyzed using big data analytical tools.

BIG DATA ANALYTICS GOAL IN HEALTHCARE

Data from all the sources in healthcare can be helpful for future insights using big data analytics which can therefore reduces costs of treatment and other expenses for payers, providers, patients and management/administration.



Source: IBM

Figure 4: Goal of Big Data Analytics in Healthcare

BIG DATA TOOLS AND TECHNIQUES

There are many tools available in market for big data analysis. Few are, as under;

1. **HADOOP** the open source software is a framework of tools from APACHE. It supports running applications on Big Data. It processes large data sets by commodity hardware. It breaks big data into proportions and sends to various commodities hardware/computers. The process is done in batches. It has two components;

- MapReduce
- HDFS

Both components contain task trackers and data nodes respectively, to send the batches for processing among various commodities hardware and recognize/manage which batch is send across to which hardware. There is a linear relationship between speed and no. of computers. Yahoo, Facebook, Amazon etc are using Hadoop.

Operating system-Windows, Linux, OS X

2. **GRIDGAIN** provide with alternate MapReduce which is still compatible with HDFS. It support in memory for fast analysis of real-time data.

Operating system-Windows, Linux, OS X

3. **HPCC** is high performance computing cluster developed by Lexis Nexis Risk Solutions. It provides superior performance to Hadoop.

Operating system- Linux

4. STORM by twitter provides real time computation capabilities. It is highly scalable, robust and fault tolerant, works with all programming languages.

OPERATING SYSTEM- Linux

There are four types of primary technologies, which are as under;

1. **GRID COMPUTING**: Central grid infrastructure that allow parallel processing for management of data, analytics and reporting.
2. **IN-DATABASE PROCESSING**: A scalable architecture offered by third-party database without getting involved with purchasing separate database, henceforth reducing time and cost.
3. **IN-MEMORY ANALYTICS**: Create and deploy analytical model quickly. In-memory access to data and do complex analytical computations.
4. **SUPPORT FOR HADOOP**: Hadoop framework using commodity hardware.

BIG DATA CHALLENGES

While big data analysis is of great help in unleashing the hidden solutions. It is a need of an hour for all industries. It has many challenges to face, which are:

- ❖ Agile information/service delivery
- ❖ Fast change in data sets, required to analyze for providing solutions
- ❖ Unpredictable business demands to be meet
- ❖ Volume and variety of data growing
- ❖ Ability to compete with growing demand

BIG DATA OPPORTUNITIES IN HEALTHCARE

Big data is being used in almost all the industries and the next opportunities lies in Healthcare. Big data can be used both from private and public hospital/health

sectors. Certain areas of healthcare opportunities are Population health management, claim management, new diseases trend analysis, fraud detection for payers, R&D, unusual patterns analysis, treatment planning, surveillance and more.

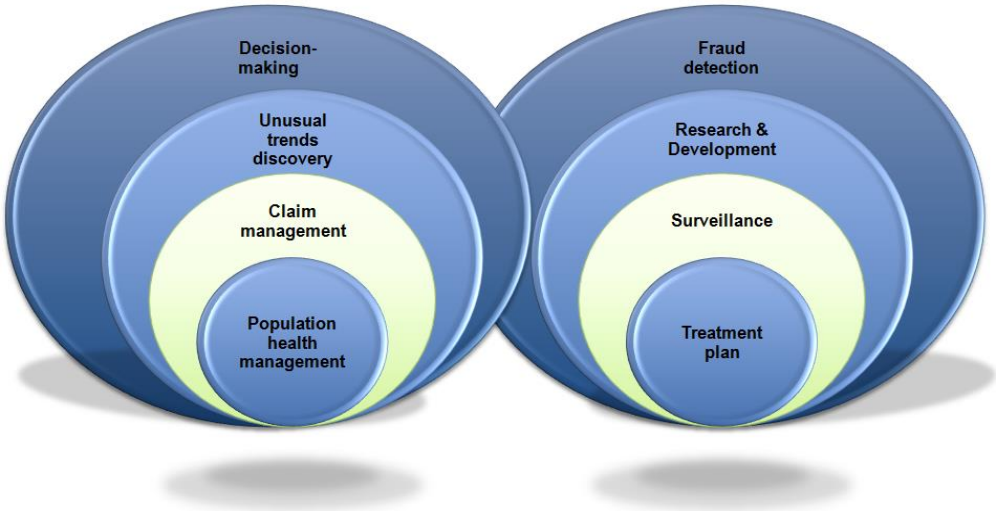


Figure 5- Big Data Opportunities in Healthcare

OPPORTUNISTIC AREAS	SCOPE
TREATMENT PLAN	Line of treatment, length of stay/treatment , etc.
SURVEILLANCE	Survey data used for medical research, disease trends, preventive management, etc.
RESEARCH & DEVELOPMENT	Screening of potential molecule for drug discovery, personalized medicine/genomic science, etc.

POPULATION HEALTH MANAGEMENT	Public health alerts, state health programming, outbreak management, etc.
CLAIM MANAGEMENT	Ease the process, save time, minimize claim rejection, etc.
TREND DISCOVERY	Changing disease trends, life expectancy trend, consumer trends, etc.
DECISION MAKING	Periodic reporting, graphic visuals, various quality measures help to optimize utilization of resources, etc.

Table no. 4- Opportunistic areas and scope of big data analysis

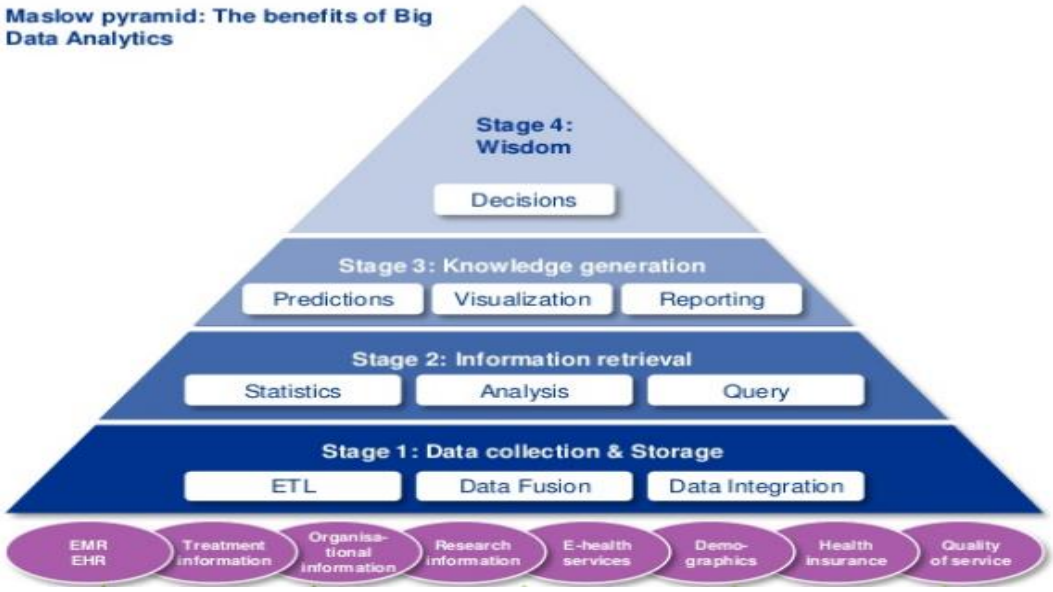
BIG DATA ANALYSIS BENEFITS IN HEALTHCARE

It has following benefits;

- ✓ It helps to improve quality of patient care delivery and tracking it's thereafter outcomes.
- ✓ Contribute in the field of drug discovery and development analysis.
- ✓ Predict future's trends by analyzing.
- ✓ Ensure health program optimization.
- ✓ Enables improved supply chain management.

- ✓ Provide solution that combine clinical, financial, and operational data to address many issues beyond inpatient care.
- ✓ It is useful for population health management.
- ✓ It predicts the scope of future business expansion. Hence transfer knowledge into improved decision making and performance.

The benefits of Big Data Analytics are visualized in Maslow pyramid by KPMG is shown below.



Source: KPMG

Figure 6: Maslow pyramid

BIG DATA ANALYSIS-PATIENT CENTRIC BENEFITS IN FUTURE

Big data analysis uses learning from past experiences and helps in deciding on:

- **LINE OF TREATMENT**: - All the relevant patient medical information and past experience from database will be used to analyze the line of treatment. It will be beneficial to decide on line of treatment for deadly diseases like HIV, Cancer, etc.
- **LENGTH OF STAY:** - Data analysis will be used to give better treatment hence reducing the length of hospital stay and treatment.
- **PERSONALIZED MEDICINE**: -Patient's past history and bio-informatics/genome analysis will be used for personalized medicine if required.
- **APPROPRIATE INVESTIGATIONS**: - Analysis will help to advice appropriate diagnostics /investigations thereby assisting fast tracking of diseases and its cure.
- **COST OF TREATMENT**: - With minimum length of stay, appropriate investigations and best line of treatment-cost of healthcare services reduces.

VARIOUS OPEN SOURCE STAKEHOLDERS

IBM's alliance with Cloud Foundry

Microsoft providing a development platform for Hadoop

Dell's Open Stack-Powered Cloud Solution

VMware and EMC partnering on Cloud

Oracle releasing its NoSql database as Open Source

SOME OF BIG DATA OPEN SOURCE SOLUTIONS

For the analysis of big data, open source solutions are available. In the below image, some of the big data open source solutions are depicted.

Source: Big data studio blog

Figure 7- Big Data Open Source Solutions

HADOOP DISTRIBUTIONS

- Hortonworks

CLOUD OPERATING SYSTEM

- Cloud Foundry — By VMware
- OpenStack — Worldwide participation and well-known companies

STORAGE

- fusion-io — Not open source, but very supportive of Open Source projects; Flash-ware applications.

DEVELOPMENT PLATFORMS AND TOOLS

- REEF — Microsoft's Hadoop development platform
- Lingual — By Concurrent
- Pattern — By Concurrent
- Python — Programming language
- Mahout — Machine learning programming language
- Impala — Cloudera
- R — MVP among statistical tools
- Storm — Stream processing by Twitter
- LucidWorks — Search, based on Apache Solr
- Giraph — Graph processing by Facebook

NOSQL DATABASES

- MongoDB-Designed to support humongous databases
- Cassandra-Used by many organization with large, active datasets, including Netflix, Twitter, Reddit, Cisco, Digg etc.
- Hbase- Non relational data store for Hadoop

SQL DATABASES

- MySql — Belongs to Oracle
- MariaDB — Partnered with SkySql
- PostgreSQL — Object Relational Database

- TokuDB — Improves RDBMS performance

SERVER OPERATING SYSTEMS

- Red Hat — The defacto OS for Hadoop Servers

BI, DATA INTEGRATION, AND ANALYTICS

- Talend
- Pentaho
- Jaspersoft

CONCLUSION

From this study various big data related concepts were understood. In the study big data definition, type of analysis, analytic processes, big data role in healthcare, its challenges, benefits, opportunistic areas etc. were discussed.

Big data is still in its infancy stage in healthcare; though it will have promising outcomes. The variety, volume, dimensions of patient e-data growing unevenly, making big data analytics in real demand.

Big data analytics, insights will be most important for getting solutions in healthcare setups. Big data will be greatly helpful for trend discoveries. It will tackle all the data sets.

Genomic study and chemical analysis for R&D purposes will be dependent on big data analytics.

RECOMMENDATIONS

After understanding various stakeholders, tools & techniques, etc., certain types were identified to help healthcare setup at this infancy stage of big data adoption, which can be set to use the available semi-structured or unstructured data for analysis. It depends on the need of an organization, kind of situation it had undergone, challenges it has been facing or targets set for future. Following are types which can be to considered before choosing any big data analytic platform;

- **CLOUD BASED DBMS ANALYTIC SOLUTIONS**: If scalability is a main challenge for you, then go for cloud based platforms-"**Do analytics in the cloud**". It provides database supports with fast processing and scalability at the same time. Also look for add on features like machine learning and statistical/graph/time series analysis .E.g. Amazon, etc.

- **SOFTWARE BASED ANALYTIC SYSTEMS**: There are software based platforms with in-built data integration, data cleansing and DBMS options. The whole process will go like: add the data, connect with platform, analyze the data and generate report .E.g. HP vertica

- **IN-MEMORY BASED DBMS SOLUTIONS**: Some vendors provide a platform with in-memory analysis using open source Hadoop. It aids the management and monitoring the workloads. If you neither have DBMS nor have will to purchase one, then this is really helpful. E.g.Altibase, Actian,etc
- **REFERENCE HARDWARE CONFIGURATION SOLUTIONS**: It helps to complement legacy data warehouse environments rather than replacing it. It provides references for Hadoop software distributors. It helps to complement Hadoop.

CASE STUDY

ASSESSING THE LEVEL OF UNDERSTANDING OF BIG DATA AND ATTITUDE TOWARD LEARNING NEW HEALTHCARE CONCEPTS AMONG HEALTHCARE IT PROFESSIONALS IN INDIA

INTRODUCTION

An easily scalable system of unstructured data with accompanying tools that can efficiently pull structured data sets is big data.-FCW Blog

Big data is an emerging challenge to manage data with large volume and dimensions, which is not in uniform format as well.

Its concept is growing day-by-day and being ready to take up in healthcare setup as well to manage pool of patient data via EHR or EMR mostly for R&D purposes. Other than R&D, it has many benefits for payers, providers and patient s.

With the adoption of EMR/EHR, all the patient data should be safeguarded using HIPAA compliance EMR/EHR. Other than HIPAA, other healthcare IT concept like HEDIS , PQRS, Meaningful use, ACO, Patient portal, Standards, ICD-10 and many more are coming across ever now and then.

Healthcare IT is the next step to achieving quality patient care delivery service via all the new concepts. Various conferences, seminars, CME etc. being introduced to physician and other professional involved in healthcare IT industry.

All the new concepts coming across are still challenge in India, starting from adoption and implementation of EMR/EHR in hospitals.

OBJECTIVE

GENERAL OBJECTIVES

- To assess level of understanding of big data and attitude toward learning of new healthcare IT concepts among Healthcare IT professionals

SPECIFIC OBJECTIVES

- To assess level of understanding of big data
- To know attitude of healthcare IT professionals toward new healthcare IT concepts.
- To encourage healthcare IT professionals, find out changes undergoing in healthcare setup in India.

METHODOLOGY

The data collection source for the study was primary data. Data collected through a structured questionnaire by interview method.

Study design	: Descriptive
Study area	: Bangalore
Study population	: Healthcare IT professionals
Sample size	: 11 (External)
	6 (Internal from organization)
Sampling method	: Convenient
Data collection	: Primary data
Tools	: Questionnaire
Data analysis	: Excel

RESULT & DISCUSSION

In this study, all the data collected through primary source is analyzed using excel as analytical tool. All the analyzed data is represented as follows;

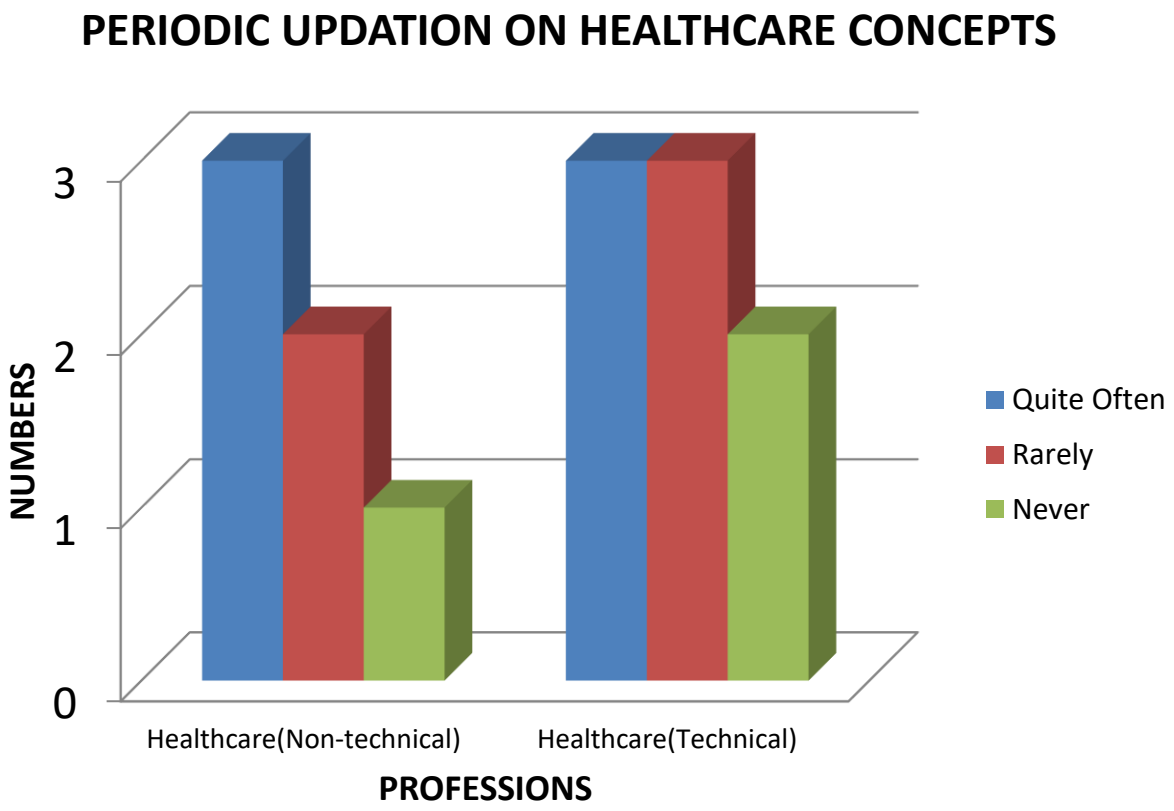


Figure 8

Percentage of respondents, update themselves on healthcare concepts;

	Healthcare(Non-technical)	Healthcare(Technical)
Quite often	24	24
Rarely	12	24
Never	6	12

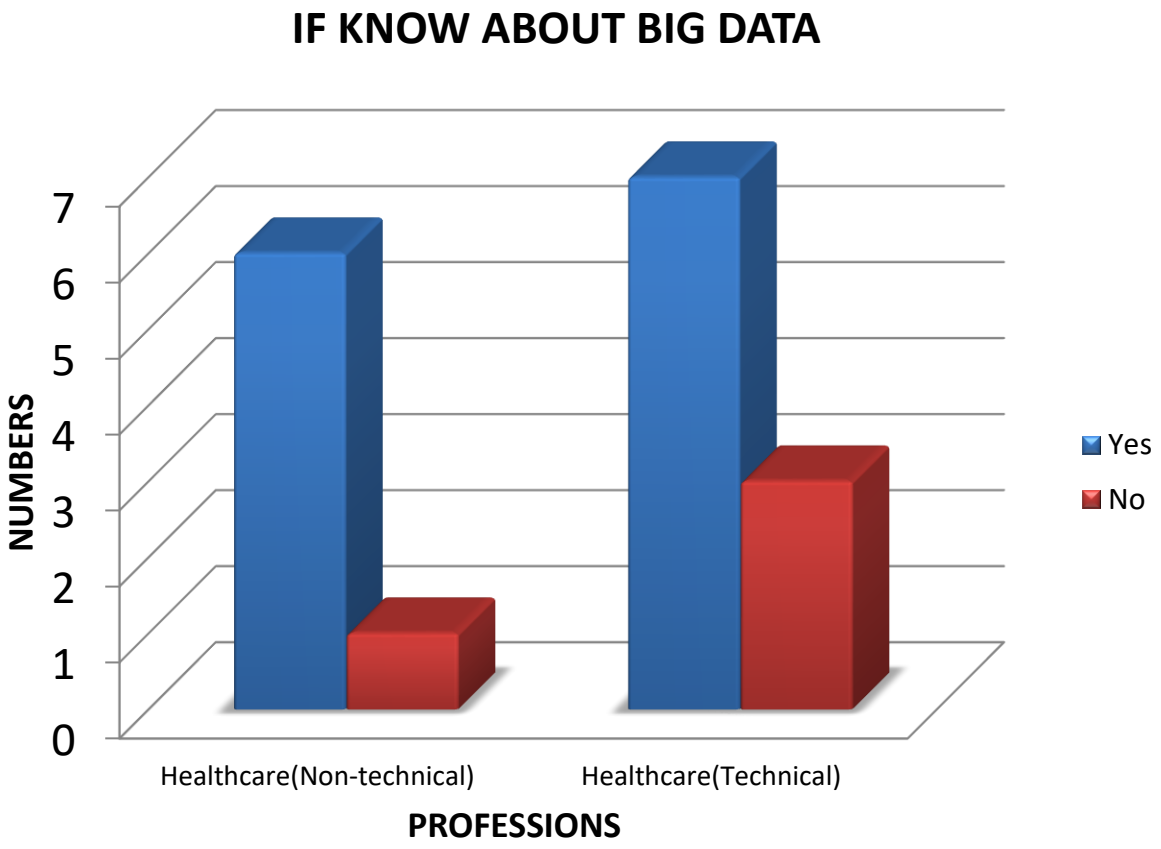


Figure 9

Percentage of respondents, know about big data;

	Healthcare(Non-technical)	Healthcare(Technical)
Yes	35	41
No	6	18

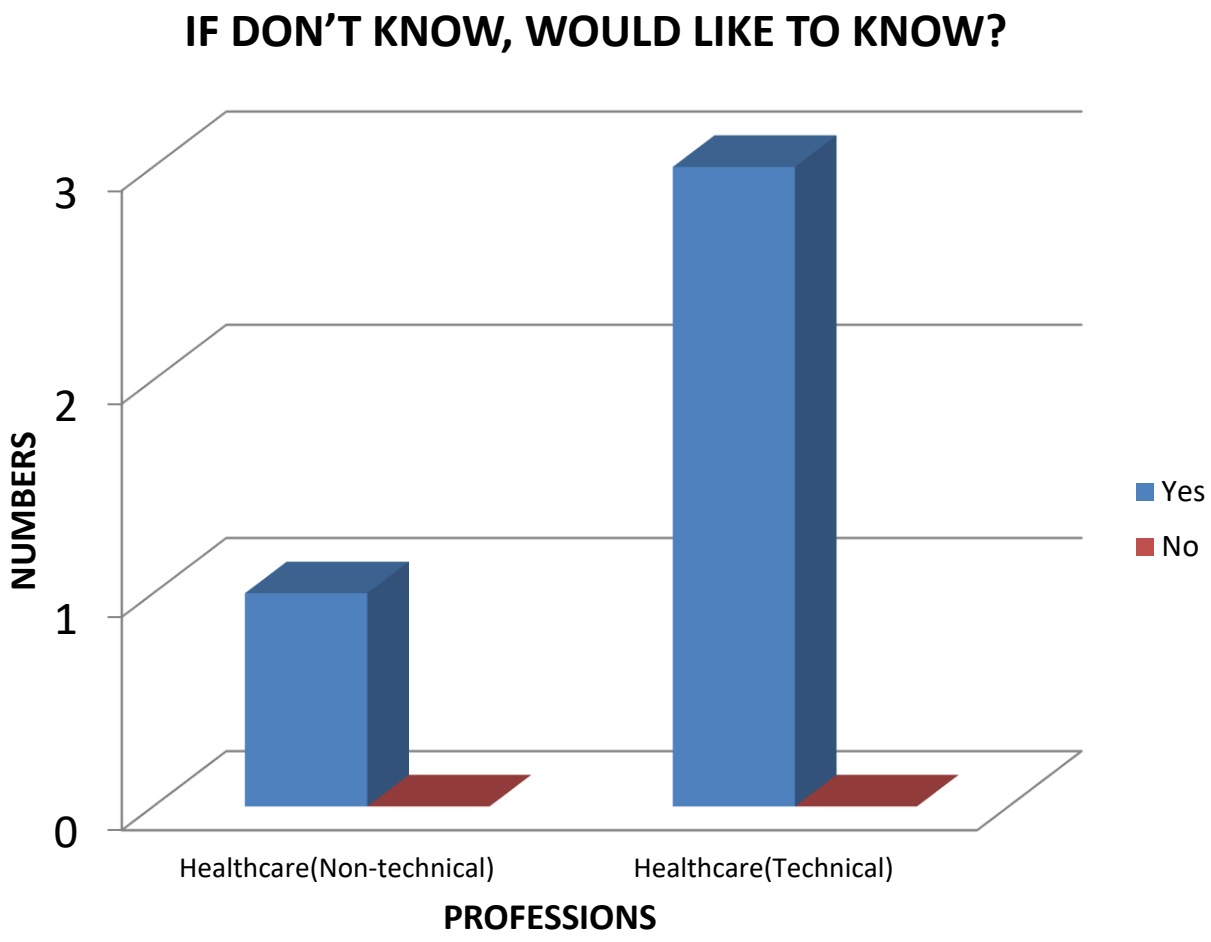


Figure 10

Percentage of respondents, would like to know about big data;

	Healthcare(Non-technical)	Healthcare(Technical)
Yes	25	75
No	0	0

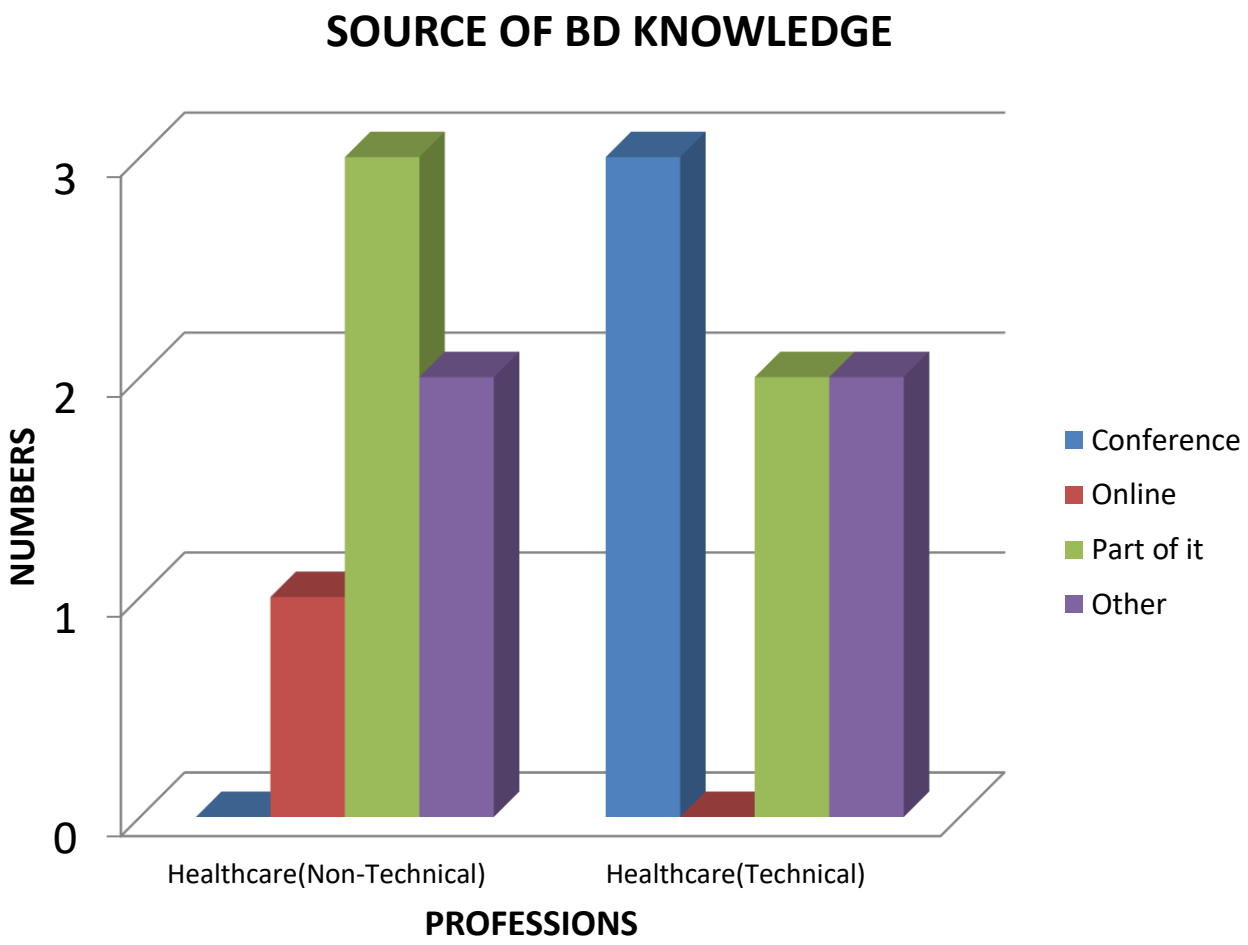


Figure 11

Percentage of respondents, have source of their knowledge on big data;

	Healthcare(Non-technical)	Healthcare(Technical)
Conference	0	23
Online	8	0
Part of it	23	15
Other	50	15

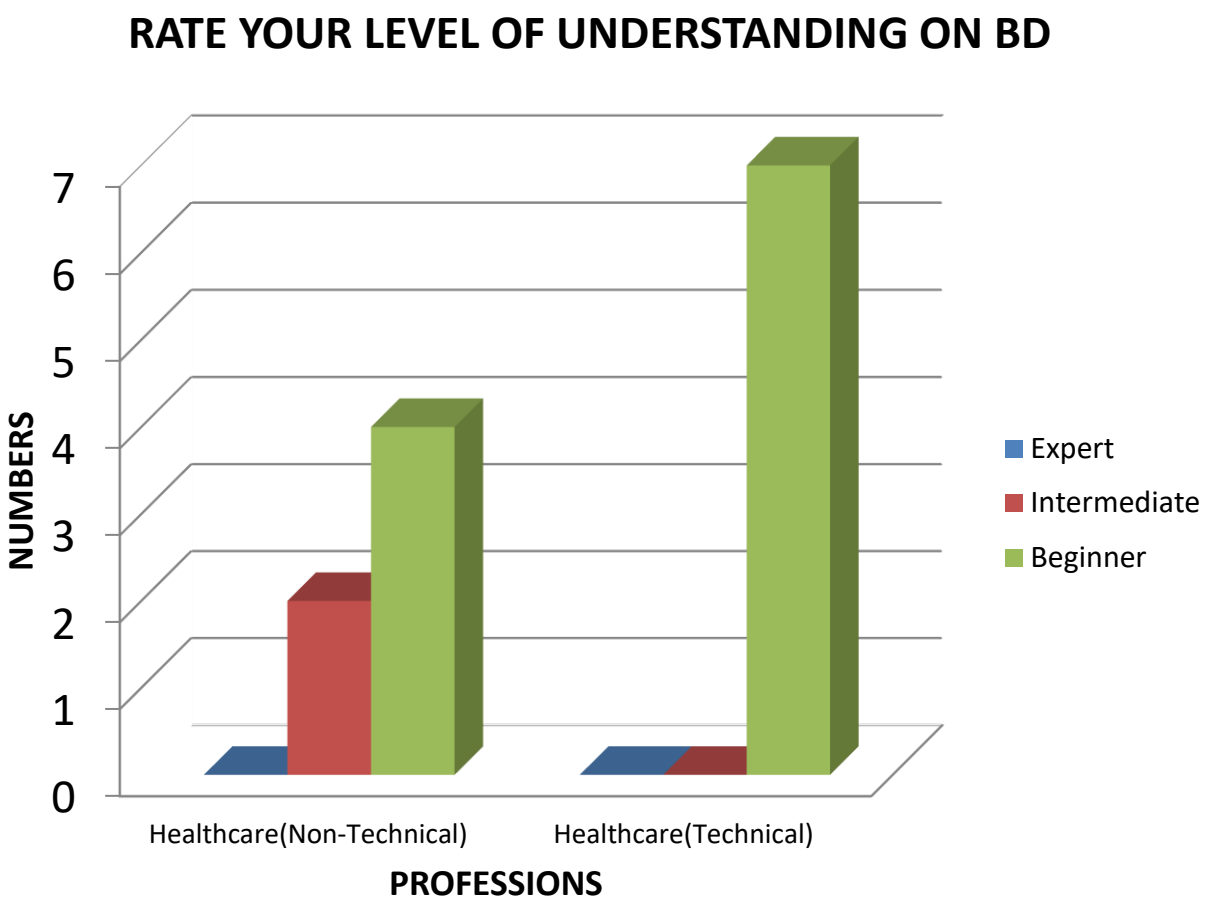


Figure 12

Percentage of respondents, rate their level of understanding;

	Healthcare(Non-technical)	Healthcare(Technical)
Expert	0	0
Intermediate	15	0
Beginner	31	54

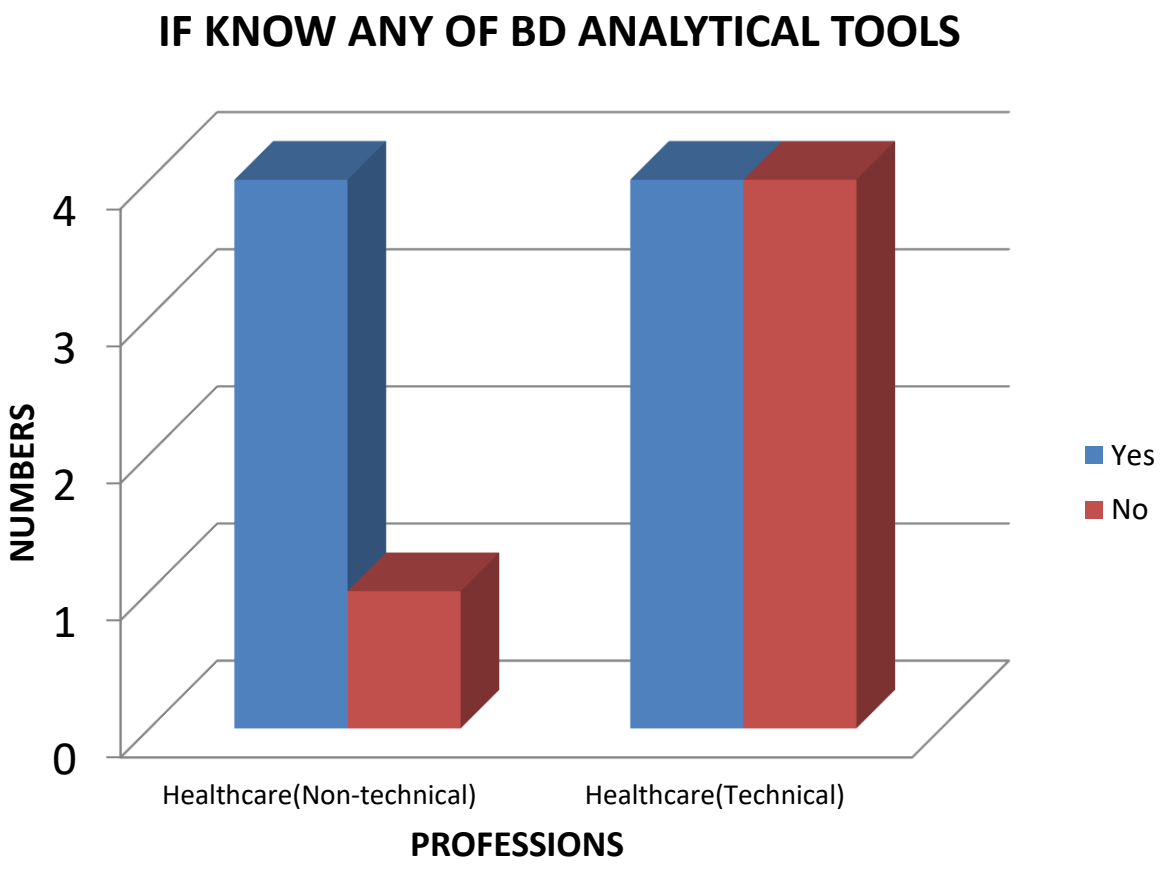


Figure 13

Percentage of respondents, if know any of BD analytical tools;

	Healthcare(Non-technical)	Healthcare(Technical)
Yes	31	31
No	8	31

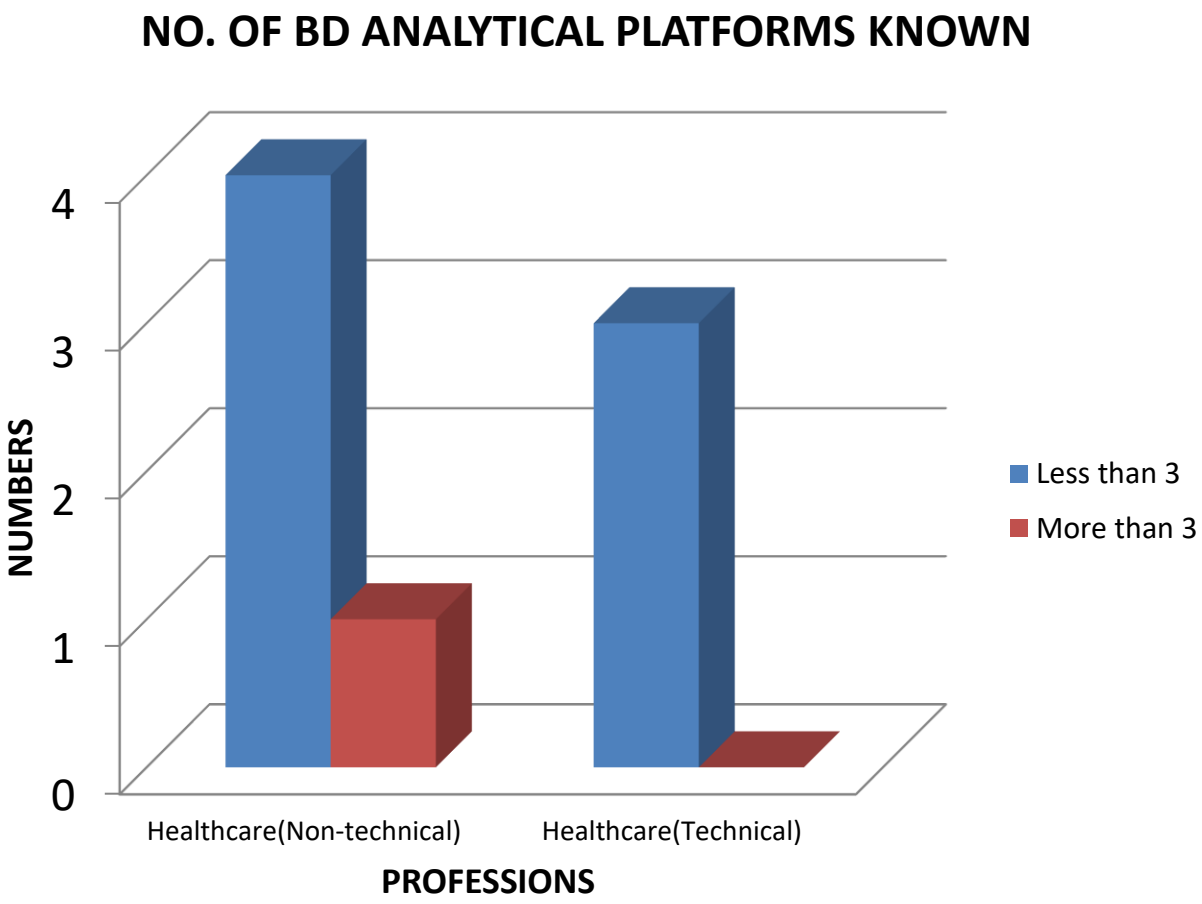


Figure 14

Percentage of respondents, know no. of platforms;

	Healthcare(Non-technical)	Healthcare(Technical)
<3	50	38
>3	13	0

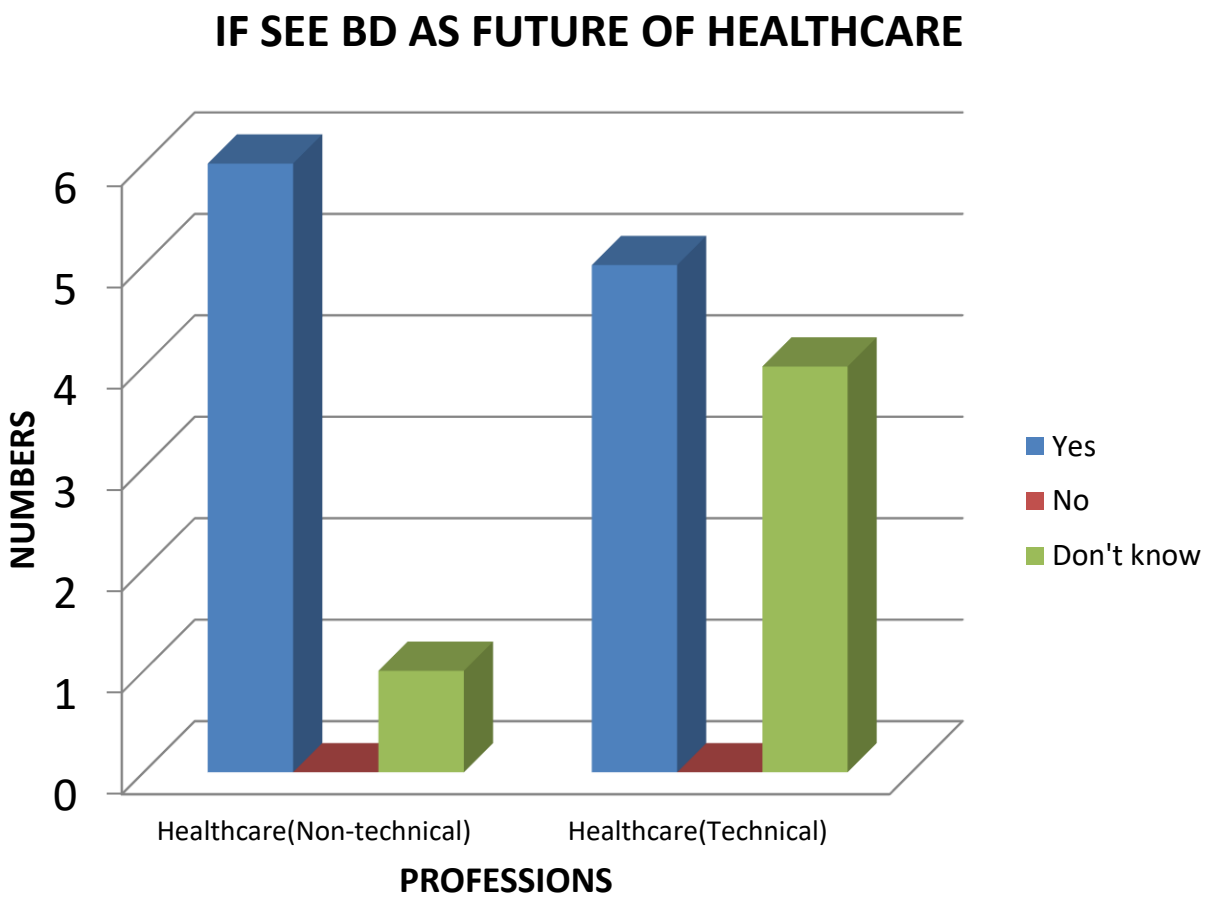


Figure 15

Percentage of respondents, see BD analysis as healthcare future;

	Healthcare(Non-technical)	Healthcare(Technical)
Yes	38	31
No	0	0
Don't know	6	25

CONCLUSION

In this case study, the sample size used was small in size to interpret anything.

All the findings, showed a mixture of responses both from healthcare technical as well as non-technical professionals. Some were aware of big data concept and know BD analytic tools while other who are unaware wanted to know about it.

Related to the attitude toward reading habits, some often kept updated on healthcare emerging topics while few never read any related journals, etc.

ANNEXURE

Questionnaire

(Tick the appropriate)

The purpose of this study is to assess the level of understanding of big data and attitude toward learning healthcare new concepts among healthcare IT professionals in India. I assure you, your identity will not be revealed in any case.

Initials: _____ **Gender:** _____

Age: _____ **Contact no. (Optional)** _____

1. Which profession are you in?
   a. Healthcare (non-technical) ☐
   b. IT (technical) ☐

2. How often do you update yourself on healthcare via journals/e-material, etc.?
   a. Rarely ☐
   b. Quite Often ☐
   c. Never ☐

3. Do you know about big data? (If yes skip question no. 4 else answer only question no. 4) ☐

a. Yes

b. No ☐

4. If no, would you like to learn about it?

a. Yes ☐

b. No ☐

5. If yes, how do you come to know?

a. Through conference ☐

b. Online material ☐

c. Part of it ☐

d. Other source ☐

6. How do you rate your knowledge on big data?

a. Expert ☐

b. Intermediate ☐

c. Beginner ☐

7. Do you know any of the big data analytical tools?

a. Yes ☐

b. No ☐

8. If yes, how many related platforms/tools do you know?

a. <3 ☐

b. >3 ☐

9. Can you see big data analysis as future, in healthcare sector?

a. Yes ☐

b. No ☐

c. Don't know ☐

10. Any comments?

……

……

……

……

……

……

*****************Thank you for your participation*********************

REFERENCE

1. (2014, May). Retrieved from My Data Mining Resources: http://mydatamine.com/what-is-the-difference-between-data-mining-and-data-warehousing/

2. Big Data. (2014, May). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Big_data

3. Frand, J. (2014, May 4). Data mining. Retrieved from http://www.anderson.ucla.edu/

4. Jee K, K. G. (2013 Jun). Potentiality of big data in the medical sector: focus on how to. PubMed.

5. Laney, D. (2001). APPLICATION DELIVERY STRATEGIES. Stamford: META Group.

6. Mark Troester. (n.d.). Big Data Meets Big Data Analytics. USA : SAS Institute Inc.

7. Ola O, S. K. (2014,Feb). The challenge of big data in public health: an opportunity for visual analytics. Online J Public Health Inform.

8. Szlezák N, E. M. (2014). The role of big data and advRanced analytics in drug discovery, development, and commercialization. Clin Pharmacol Ther , 492.

9. Bertolucci, J. (2013, June). Hadoop: From Experiment To Leading Big Data Platform. Retrieved 2014, from InformationWeek:

http://www.informationweek.com/software/hadoop-from-experiment-to-leading-big-data-platform/d/d-id/1110491

10. Leading Priorities for Big Data for Business and IT. eMarketer. October 2013. Retrieved January 2014

11. Data Mart. (n.d.). Retrieved May 2014, from Oracle: http://docs.oracle.com/html/E10312_01/dm_concepts.htm

12. Elena Kvochko, Four Ways To talk About Big Data (Information Communication Technologies for Development Series). worldbank. org. Retrieved 2014-04-30

13. Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved May 2014.

14. Data, data everywhere. The Economist. 25 February 2010. Retrieved April 2014.

15. Berger ML, D. V. (Mar,2014). Big data, advanced analytics and the future of comparative. PubMed .

16. Vie LL, G. K. (2013). The Person-Event Data. PubMed .

17. DBMS - database management system. (n.d.). Retrieved April 2014, from Webopedia: http://www.webopedia.com/TERM/D/database_management_system_DBMS.html

18. (2013, Feb). Retrieved March 2014, from Aguai solutions: http://www.aguaisolutions.com/

19.MeiLing's BI E-Portfolio. (2009, Nov.). Retrieved from WordPress: http://meilingbieportfolio.wordpress.com/category/week4-data-warehouse-and-olap/

20. Ferguson, M. (Oct,2012). Architecting A Big Data Platfom for Analytics. IBM.

21. Konkel, F. (April,2013). TheConversation. FCW.

22. O'Brien, T. (2014, Jan. 22). Data Science Report. Retrieved from OPEN SOURCE TOOLS FOR BIG DATA: http://blog.starbridgepartners.com/2014/01/22/50-open-source-tools-for-big-data-see-anything-missing/#.U3JDi_mSybd

23. grzegorzglab. (2013). Retrieved from Big-Data regula 4V: http://grzegorz.glab.pl/

24. Jimeng Sun, C. k. (n.d.). Big Data Analytics for Healthcare. Retrieved from IBM: dmkd.cs.waye.edu