Dissertation

On

"BIG DATA ANLYTICS IN HEALTHCARE"

Submitted By

Maninder Singh

PG/15/044

Under The Guidance Of

Prof. Nishikant Bele

Post Graduate Diploma in Hospital and Health Management

2015-2017

INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT & RESEARCH

Contents

Acknowledgement

Hard work, guidance and perseverance are the pre requisite for achieving success. Support from an enlightening source helps us to proceed on the path to it. I wish to thank first of all the almighty that provided me energy for the successful completion of internship.

I am thankful and obliged to the Senior Manager of ***NTT Data services*** – ***Mr. Ajay Aiyar*** and IT Services Managers- ***Ms. Archika Roy, Ms. Avishikta Sarkar***, ***Mr. Venkateshwarrao M*** and ***Mr. Rituraj Chaudhary*** for giving me an opportunity to work on this project. I am also thankful to my mentor ***Mr. Hitesh Mehta, Dr. Sushma Baradwad*** and ***Mr. Gourav Tomar*** my trainer ***Mr. Yatharth Sharma***, ***Ms. Gabreena Karishma Lewis*** and whole HPF Team for their continuous support, guidance and perseverance during the course of my project.

I am highly indebted to my college mentor ***Prof. NishiKant Bele*** & ***Prof. Aanandi Ramachandran*** for valuable guidance and motivation on various aspects of project.

I would also like to acknowledge ***Dr. Ashok K Agarwal*** Dean of International Institute of Health Management and Research, Delhi for his support.

Furthermore, I would like to thank my dear classmates & friends ***Dhruvika Kohli, Neelam Soni, Diksha Sharma, Aayushi Soni, Ritwik Chawla, Manisha Raturi, Sachin Kumar, Asib khan, Anita Verma*** who have supported me throughout by keeping me encouraged and helping me putting pieces together.

In a special way I would like to thank my parents ***Mr. Jasbir Singh*** and ***Mrs. Vimla Singh*** and my sisters ***Ms. Sudha & Geeta*** who have always been there for me, with their unconditional love, support and faith in me throughout my life and helping me complete my study.

Last but not least, I would like to extend my sincere thanks and deep gratitude to all those who have been my guiding force in carrying out this piece of work.

It has been my good fortune to be benefited by their knowledge, deep insight without which this project would not have taken the exact shape .To them, I tender my heartfelt regards.

Thank you

Maninder Singh

List of Figures

Abbreviations

1. US: United State
2. GPS: Global Positioning System
3. DVD: Digital Versatile Disc
4. EKG: Electrocardiogram
5. ICU: Intensive Care Unit
6. PDF: Portable Document Format
7. HMO: Health Management Organization
8. EMR: Electronic Medical Records
9. EHR: Electronic Health Records
10.CDR: Clinical Data Repository
11.ICD: International Classification of Diseases
12.HDFS: The Hadoop distributed file system
13.POSIX: The Portable Operating System Interface
14.CPU: Central processing unit
15.SQL: Structured query language
16.XML: Extensible Markup Language
17.JAQL: Jason Query language
18.API: Application program interface
19.DNS: Domain Name System
20.HTTP: Hyper Text Transfer Protocol
21.URL: Uniform Resource Locator
22.JSON: JavaScript Object Notation
23.CDSS: Clinical Decision Support System
24.BDA: Big Data Analytics
25.ECL: Extensible Computer Language

Organization Profile

NTT DATA is a top 10 global business and IT services provider and global innovation partner with 100,000+ professionals in more than 50 countries now with $16B in revenue.

Headquartered in Tokyo, NTT DATA puts emphasis on long-term commitment and combine global reach and local intimacy to provide premier professional services from consulting, system development to business IT outsourcing. Since 1967, NTT DATA has played an instrumental role in establishing and advancing IT infrastructure. Originally part of Nippon Telegraph and Telephone Public Corporation, its heritage contributed to social benefits with a quality-first mindset. A public company since 1995, the company builds on this proven track record of innovation by providing novel IT solutions to bring results in greater quality of life for people, communities and societies around the world.

Who we serve

- Industry-specific consulting
- Digital business services
- BPO and BPaaS
- Business intelligence, analytics and automation
- Enterprise applications and SaaS
- Application modernization, development and management
- Cloud services
- Infrastructure management, security and hosting

Who we serve

- 50+ federal agencies and military branches, 25 states and municipalities
- 50% of U.S. hospitals, top 10 health plans, millions of covered patients
- Top 25 leading financial institutions in North America
- Top 10 automotive companies worldwide
- Manufacturing customers in over 40 countries

Who we are

- 100,000+ professionals in 50+ countries
- 6,000 research professionals and dedicated R&D facilities
- 9,000 SAP professionals
- 10,000 financial services and insurance specialists
- 15,000 skilled manufacturing resources
- Trusted U.S. government partner for 50+ years

What we do

- 10th Largest in IT Services Worldwide in 2015 by Market Share* (NTT DATA)
- 7th Largest in Global Implementation Services in 2015 by Market Share Worldwide+ (NTT DATA)
- A Leader in The Forrester Wave™: North American Workplace Services, Q4 2015^ (Dell Services)
- Leader in the NelsonHall Digital Transformation Services NEAT Evaluation^* (Dell Services)
- 12th Largest in Consulting Services in 2015 by Market Share Worldwide- (NTT DATA)

NTT Group consists of major companies like Nippon Telegraph and Telephone Corporation, NTT Communications Corporation, Dimension Data plc, NTT DOCOMO, INC. and many subsidiaries all over the world. Taking advantage of this opportunity of this scale, NTT DATA achieved a number of significant successes by collaborating with NTT Group and it provided enormous creative synergy.

The goal of NTT has been to create a foundation for future business by incorporating a number of overseas companies in order to establish a framework through which we can provide our diverse services, as typical Japanese courteous service, worldwide to support our customers' needs. As one of the global innovators, NTT are always challenging more innovative business approach and enhancing our creativity by respecting diversity.

John W. McCain is the Chief Executive Officer of NTT DATA Services headquartered in Dallas, Texas, USA. He is a member of the NTT Holdings Global Strategy Committee and serves as senior vice president of NTT DATA Corporation.

Dan Allison is the President, Global Healthcare and Life Sciences. As head of the company's largest industry segment, Dan is responsible for leading the growth, profitability and transformation of the global healthcare business, which focuses on provider, physician, health plan and life sciences clients. Dan has more than 30 years of leadership experience in IT outsourcing and business process outsourcing services in various verticals, with a strong focus in healthcare.

Americas

North America

In North America, NTTDATA partnered with a range of businesses and government agencies providing a flexible array of engagement options, including consulting, managed services, outsourcing, and the cloud.

Leveraging strong technical know-how, practical industry insights, and global reach, it relentlessly drives improvement across systems and processes while increasing business flexibility. The company is focused on getting faster results with less risk, so its clients can flex their businesses to respond to changing market dynamics and capitalize on growth opportunities.

Latin America

NTT DATA entered the Latin American market through the acquisition of the Value Team Group, a specialist in IT consulting and services. Today, the company provides a wide offering of customized services and end-to-end solutions. The aim is to enable customers to grow and stand out from the competition by adopting innovative IT concepts and technologies.

Europe and Middle East

Over the past few years, we have expanded our IT service networks in Europe through the acquisition of a majority stake in intelligence, Cirquent, Value Team, Intelligroup and Keane.

NTT DATA Group offers best-in-class consulting services and enterprise solutions for industries in the manufacturing, banking, insurance, telecommunications, media, energy, retail, service and public sectors. Our consulting services range from business process consulting to conceptual design, implementation and integration, as well as the support, operation and maintenance of IT systems. Additional offerings include outsourcing, hosting and full-service solutions in the ERP environment.

APAC/ India

NTT DATA positions APAC and India region as both an emerging market and the delivery resource pool to provide cost competitive and high quality service in our global strategy. The company address both multinational corporations and local client in this region. With global capabilities, NTT DATA support multinational corporations, primarily in Healthcare, insurance, automotive and electronics industries in rapidly growing APAC market. In addition, NTT DATA offer the services to local clients in both financial and public sector by leveraging our accumulated experience across the world.

NTT in Healthcare:

Healthcare companies are balancing the quality and cost of care while serving a rapidly aging population and rising healthcare costs. At the same time, those firms are facing escalating competition, the feared patent cliff for many blockbuster drugs, and changing regulations and standards.

NTT DATA partners with some of the world's leading healthcare organizations to help them proactively manage their business through the use of information, data, and technology. In fact, its technology-enabled services support over thousands of organizations within the sector, enabling them to rapidly and cost-effectively adjust to dynamic market and regulatory demands.

Industry Recognition

- Positioned by Gartner in the “Leaders” quadrant of the Gartner Magic Quadrant for Data Center Outsourcing and Infrastructure Utility Services, North America for the fifth consecutive year.

- Ranked “#1 IT Services Provider to Healthcare Providers,” by Gartner for the sixth straight year.

- Positioned as a leader in Everest Group’s “IT Outsourcing in the Healthcare Provider Industry—Service Provider Landscape with PEAK Matrix Assessment” for a third consecutive year.

1. Abstract

Big data is itself a vast concept. In today's scenario big data plays a very important role in different fields and business, by using different tools and techniques, analysts show the different trends and results for business benefits and it analyzes the future trends of markets. Big data analytics has two distinct supposition – one is big data and second is analytics. By meeting this, it serves as a new information management approach that has been designed to derive previously untapped intelligence and insights from data to address many new and important questions. Within the healthcare, big data provides shareholders and collaborators with new insights that have the potential to advance personalized care, improve patient outcomes and avoid unnecessary costs.

This report defines the big data analytics, its potential, architecture, characteristics and frames and its economic value, challenges and benefits in the healthcare.

Objectives: To describe the potential of big data analytics in healthcare. To analyze the challenges with big data analytics in healthcare. To find out the various benefits associated with big data analytics in health care.

Methodology: This study is based on a systematic literature review of Big Data and Big Data Analytics and their relation to improvement the healthcare in patient health and other. Biomed, Google Scholar, Scopus and Wikipedia and Google Search were searched in March and April 2017, and PubMed and biomed and google for search engine for articles that discussed the Big Data analytics in Healthcare and Tools and techniques used in Healthcare.

Results: Big data analytics has offered a brand new way to develop actionable insight, organize their future vision, maximize the outcome and scale back time to worth. This approach is additionally useful to predict perceptive info to the healthcare enterprise regarding their management, planning and the measurement. The evaluate result can enhance or help to enhance the decision making capacity of the top management.

Conclusion: Big data analytics has the potential to remodel the means of life sciences and health organization use subtle technologies to achieve insight from their clinical and alternative data repositories to form conversant selections. Analytics enable organization to analyze and explore data to spot relationships, trends and patterns, to reveal insight that. Once combine with the business context, create information, within the future, the implementation and usage of massive data can unfold speedily.

2. Introduction

The phrase “Big Data" has launched a veritable industry of processes, technology and personal to support what appears to be an exploding new field. Big companies like Amazon and Wal-Mart and NASA are using Big Data to meet their business and/or strategic objectives. Big Data also play a role for small or medium-sized companies and organizations that recognize the possibilities to capitalize upon the gains. Big data is the current buzzword and by all means, it is going to affect healthcare.

In the present healthcare business environment, healthcare providers should understand that big data analytics is a necessity and not a luxury and so is the understanding of big data analytics. Big data analytics has many potential to impact healthcare positively in which first is improving quality of care, saving lives and lowering costs. “Fundamentally, big data is helping organizations become more productive, efficient and reduce costs. Like many other industries, healthcare has adapted to data analytics for improving patients’ quality of life, not only for its financial returns” Indian providers have been using electronic health records (EHR) and hospital information systems (HIS) to make their organization productive and profitable.

These technologies in our day to day life collectively generate a huge numbers of data. Indian Healthcare industry gadgets, insurance claim, diagnostic tests, prescriptions, patient care records and medical research are continuously generating zettabyte of data. However, ‘big’ in big data analytics not only defining the size but also defining the intricacy and quality of data. “We can explain big data could be as the total comprehensive data about an entity encompassing all sources. For better understanding of big data, it is very important to understand that what type of data it is and how it is different from information, which are quite often thought of as one and maybe same,

Data should be considered as raw information, filter and without filter, duplication, or create the forms that are the building block for information. Data can be transformed to information only when some logic can be added to present a particular fact or view point Information gathered from big data is often more structured, substantial and unique, hence its value.

The definition of Big Data and Big Data analytics is:

Big Data: There is nothing new about the big data which has been around since at least 2001, in the nutshell, big data is phrased used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In enterprise scenarios the volume of the data Is too big or its move too fast or it exceed current processing capacity it define the four v's and the four v's – volume, velocity, Variety and veracity.[1]

Volume	Velocity	Variety	Veracity

Big Data Analytics: *"Big data analytics is the advanced analytic techniques to analyze very large data sets that include different types such as structured/unstructured. Analyzing big data allows business users and researchers to make better and faster decisions using data that are inaccessible. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses and researchers can analyze previously untapped data to gain new insights with better and faster decisions."* [2]

There are many sources of data in healthcare and it comes from many sources like transactional data machine to machine data, biometric data, human generated data, as well as web and social

media data. This data has to be polled and cleansed and readied for the purpose of big data analytics. “As big data is all the data about an entity, say for example healthcare, the existence of it is distributed across multiple sources and hence kept in a distributed fashion all over. The data may not be all electronic or digital either. Hence, it is likely it will not be under any specific control either and it will have multiple ownerships. Maximum people are using social media and it is so popular today and lots of data would be on social sites and if one could get to them, a lot of intelligence could be harnessed out of it.

In the developed countries, national registries and state departments collect healthcare data and composite it over the years. Thus this big data is available for scientist to work on it. However, in India healthcare records are stored and generated by individual health organizations. There is a possibility that this data may be in redundant and is difficult to access. Having said that, few health organization, for the benefits of the patients, they agree to share the data but even it is a very difficult task to get all similar data on one platform.

According to a report ‘Big Data Vendor Revenue and Market Forecast 2011-2026’ by Wikibon the US Big Data market reached $27.36 billion in 2014 and is slated to grow to $84 billion in 2026. One of the factors driving growth of the big data market was the increasing establishment of big data-driven decision making as a key strategic priority in board rooms and C-suites across vertical markets but particularly in the financial services, retail, healthcare and telecommunications industries.

In this review first we will understand the big data and use of big data in healthcare. This will explore the how big data can be applied to a particular area through which we can get benefits for the target research.

3. Objectives

1. To describe the potential of big data analytics in healthcare.
2. To analyze the challenges with big data analytics in healthcare.
3. To find out the various benefits associated with big data analytics in health care.

3.1 Research Questions:

1. What is big data and architecture and tools and techniques associated with big data?

2. What is the potential of big data analytics in healthcare?

3. What are the various challenges with big data analytics in healthcare?

4. What are the various benefits associated with big data analytics in healthcare?

4. Review of literature

4.1 Big Data:

We are humans and we are creating data every day and we create 2.5 quintals (bytes) of data every day, it is so much 90% of data created in the world last few year. The source of this data is everywhere and data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

Big data- large volume of data –it is both structured and unstructured. It is being increasingly used everywhere in planet. Big Data has the potential to help companies improve operations and make faster, more intelligent decisions. This data, when captured, formatted, manipulated, stored, and analyzed can help a company to gain useful insight to increase revenues, get or retain customers, and improve operations. Big data include it define the four v's and the four v's – volume, velocity, Variety and veracity, refer fig. 3.1 & 3.2.

Volume:

Volume in big data defines the enormous amount of data that are created by humans. But now the data is generated from networks machines and human interaction on system for example social media the volume of data is massive. Volume of data is not much as the problem.

"*It is being observed the volume of global data increasing exponentially, from 130 Exabyte's (an Exabyte is* 10^{18} *bytes of data) in 2OO5 to 7,910 Exabyte's in 2015.*

It is predictable that by 2020, 35 zettabytes (10^{21} *bytes) of digital data will be there —a stack of DVD's that would reach halfway from the Earth to Mars.*

However, it been observed that 20% of the world's data is structured (which is suitable for computer processing), with lots of unstructured data growing at 15 times the rate of structured data. [4] It has observed that in next 3 years more than 1 billion smartphones will enter service, 400 million new tablets will connect to the Internet and there will be 1 billion active personal computers in the world. [3]

Velocity:

Velocity deals with the stride at which data flows in form sources like machine, networks, business processes, and human interaction with thing like mobile devices and social media etc. This flow of data is continuous and massive and the real time data can help researchers and businesses to take valuable decision that provide strategic competitive advantages and release of investment. If you are able to handle the velocity that data can help deal with the issues like velocity and volumes.

"Most healthcare data has traditionally been physical —, X-ray films, paper files, scrips. But in some healthcare setups and in medical situations, real-time data for ex. trauma monitoring for blood pressure, operating room monitors for anesthesia, bedside heart monitors, etc. it became a matter of life or death. In between are the medium-velocity data of multiple daily diabetic glucose measurements (or more continuous control by insulin pumps), blood pressure readings, and EKGs.

In future, few applications of real-time data in the ICU, for example detecting infections as early as possible, it will identify swiftly and apply the right treatments, could reduce patient morbidity and mortality or even stop hospital outbreaks. Real-time streaming data can already monitor neonates in the ICU, to predict life-threatening infections sooner." [4]

Veracity:

In big data, veracity refers to the noise, biases, and abnormality of data. This data is being stored and mined meaningful to the problem being analyzed. Veracity in data is a big challenge when compare two things like volume and velocity.

Data quality is a particular concern in healthcare because of two reasons, of which one is the right health information because it is a matter of life and death and the other includes the quality of healthcare data since it is often unstructured, unreadable prescription and incorrect.

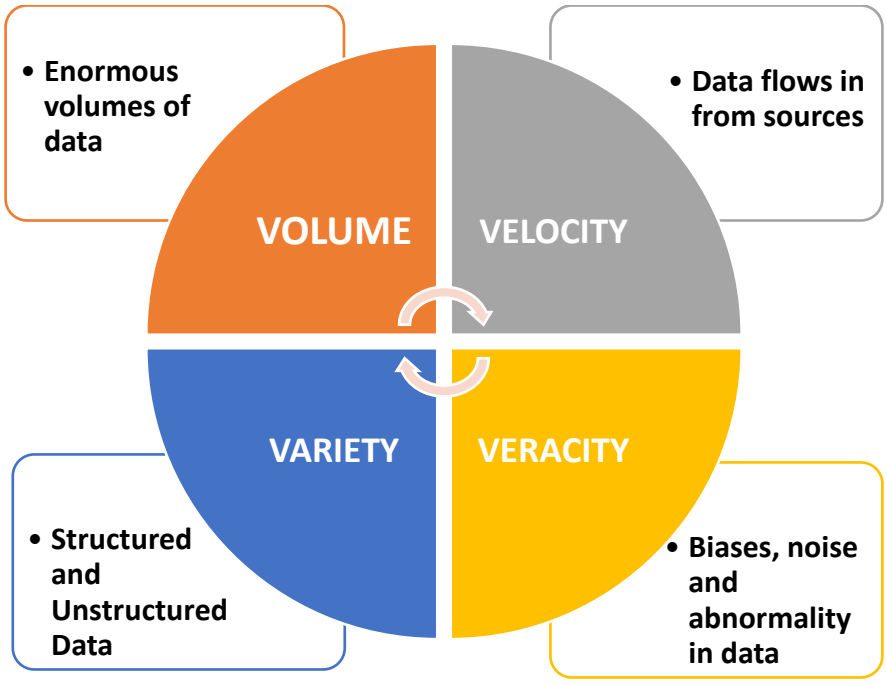


Fig.3.1 The four v's in Big Data

Variety:

Variety in big data refers to the many sources and types of data. It could be structured and unstructured. We used to store data from sources like spread sheet and database. In today's scenario data comes in the form of photos, emails, videos, monitoring devices, audios, PDFs, audios, etc. This type of unstructured data creates lots of problem for storage, mining, and analyzing the data.

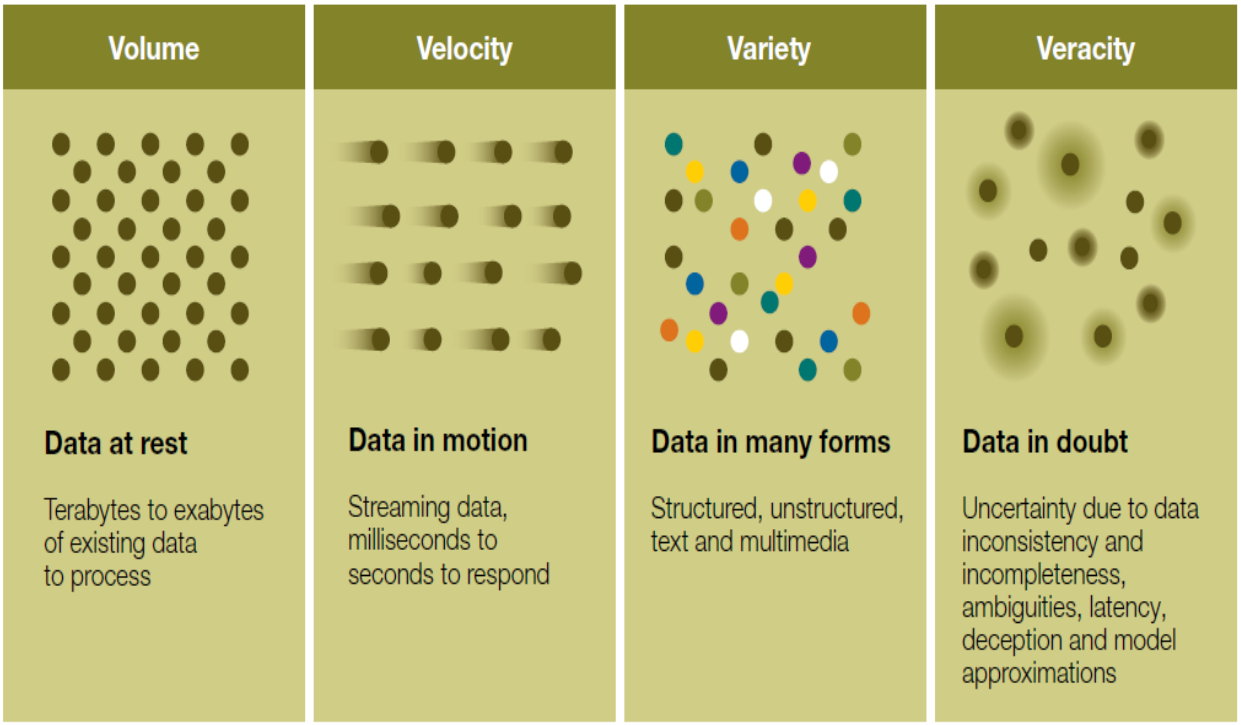


Fig.3.2 The four V's in Big Data

4.2 Big Data Architecture:

The traditional health informatics and analytics projects are similar to the conceptual frame work for a big data analytics projects in healthcare. Only the key difference is how data is being processed. In a regular healthcare project, the analysis can be performed by the analyst with a business intelligence tool installed on a stand- alone system. Big data by definition is large, and processing is broken down and executed across multiple nodes. The distributed processing concept has existed for decades. It is used for analyzing large volumes and sets of data, as healthcare provider have to tap into their large sets of data repositories to gain insight for making good informed decision related to healthcare. Moreover, Hadoop/MapReduce and other are open source platform available on cloud, they have encouraged the big data analytics applications in healthcare. While the models and algorithm are similar, traditional analytics tools and its user interface, used in big data are completely different; these analytical tools have become super user friendly and transparent. Big data analytics tools, on the contrary, are immensely complex, programing intensive, and required of a variety of skills.

All are emerged in an ad hoc fashion mostly as open source platform and tools, that's why they deficit the support and user friendliness that vendor- driven prosperity tools process. Complexity begin with data itself. Healthcare big data can come from external sources as well as internal sources, example of internal sources are, electronic health records, clinical decision support system, etc. and examples of external sources are laboratories, government sources, pharmacies, insurance companies & HMO etc.

Sources And Data Types Include:

Few examples of data's are:

- **Web and social media data**: lots of interaction data from Twitter, Facebook, LinkedIn, blogs, and the like with this it also include health plan web sites, smartphones app.
- **Machine to machine data:** data generate from remote sensors, gadget, meters, and other vital sign devices.
- **Big transaction data**: health insurance claims and other billing records increasingly available in a formats of semi-structured and unstructured formats.
- **Biometric data**: x-ray, genetics, handwriting, finger prints, retinal scans, x-ray and other medical images, blood pressure, pulse and pulse-oximetry readings, and other similar types of data.
- **Human-generated data:** Human generate both unstructured and semi-structured data and few examples of unstructured and unstructured data is EMRs, physicians Notes, email, and paper documents.

Other source and data type I have already explained above. For the purpose of big data analytics, data should be pooled. Data in raw state need to be processed or transformed. In a service oriented architectural approach combined with web services is one possibility. The data become raw and services are used to call, process and retrieve the data. Data warehousing is another approach for aggregation and processing of data, if data is not available on real time. The steps of extract transform and load. (ETL) Data from diverse sources is cleansed and readied.

Depending on data is unstructured or structured, several data formats can be input to the big data formats can be input to the big data analytics platform. In the conceptual framework, several decision are made regarding the tool selection, the data input approach, distributed design, and analytics models.

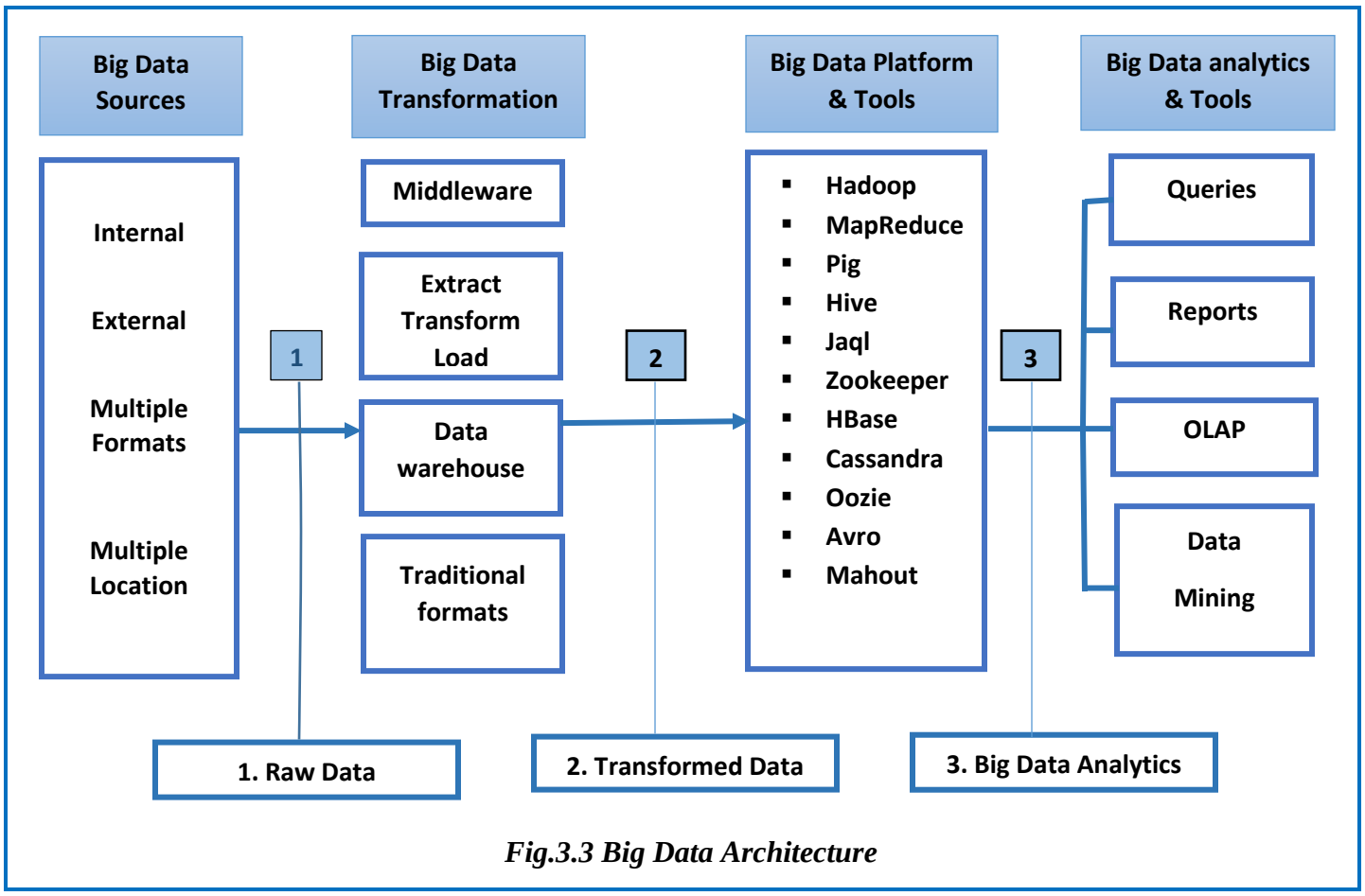


Fig.3.3 Big Data Architecture

So main four application of big data used in healthcare and these include OLAP, queries, data mining, reports. Across the four application, visualization is an overreaching theme. A wide verity of technique and technologies has been developed for statistics, computer science, applied mathematics and economics for manipulate, aggregate, analyze, and visualization of big data in health care.

Hadoop (Apache Platform) is the most significant open source distributed data processing platform, earlier developed for aggregating web search indexes. It belongs to "NoSQL" technologies and others includes MongoDB and CouchDB – they evolve to aggregate data in unique ways.

Hadoop is a tool and it has potential to process tremendously amount of data by allocating data sets to the numerous data servers and each of which server solve larger the problems of different parts and integrates them for the final results. Hadoop plays double role of data organizer and analytics tool and it offer the potential in enabling enterprise to mobilize the data that has been difficult to manage and analyze. Specifically, Hadoop make it possible to process hugely large volume of data with various structured and unstructured at all. Installation, configuration, and administration of Hadoop can be challenging because Hadoop skills are not easily found. For these reason organizations are not quite ready to embrace Hadoop completely. There are various platform and tools support the Hadoop distributed platform. Numerous vendors like, Cloudera, Hortonworks and MapR technologies distribute open source Hadoop platform, other proprietary option are also available such as IBM, Biginsight and further, maximum of these platform are available cloud version, which making them wildly available. HBase, and MongoDB, Cassandra described above, are used widely for the database component. While the available framework and tools are usually open sources and rapped around. While the available frameworks and tools are mostly open source and wrapped around Hadoop and related platforms, there are various trade-offs that developers and users of big data analytics in healthcare must consider. While the development costs may be lower since these tools are open source and free of charge, the downsides are the lack of technical support and minimal security.

In the healthcare industry, these are significant drawbacks, and therefore the trade- offs must be addressed. Theses platform required great deal of programming, skills the typical end user in healthcare may not process. Furthermore, considering the sole emergence of big data analytics in healthcare. Privacy ownership issue raise by government, security and standard have yet to be addressed. [5] [6] [7]

4.3 Sources of Big Data in Healthcare:

According to an article of "sources of big data in medicine" Big Data in healthcare is defined as the data totally coming through healthcare industry and wellbeing. There is broad view and sources and types of big data in healthcare researcher, payer, policymaker and industries, refer fig. 4.1 & 4.2.

Clinical information system: Clinical information system is computer based system that is that is designed in a way for collecting, storing and manipulating and making available information important to the healthcare delivery process. These are the traditional sources of clinical data that health care provider are using.

- **Electronic Health Record (EHR's):** An Electronic Health Record (EHR) is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that persons care under a particular provider, including:

  - Demographics
  - Progress notes
  - Problems
  - Medications
  - Vital signs

- Past medical History
- Immunizations
- laboratory Data
- Radiology Reports

The EHR automates access to information and has the potential to streamline the clinician's workflow. The EHR also has the ability to support other care-related activities directly or indirectly through various interfaces, including evidence-based decision support, quality management, and outcomes reporting. [8]

- **Electronic Medical Records (EMR'S):** An electronic medical record (EMR) is a digital version of the traditional paper-based medical record for an individual. The EMR represents a medical record within a single facility, such as a doctor's office or a clinic. [9]

  Components of electronic medical records

  - Demographics Details
  - Progress notes
  - Nursing notes
  - Laboratory reports
  - Radiology reports
  - Operative records etc.

- **Health Information Exchange (HIE):** Health information exchange (HIE) is the mobilization of health care information electronically across organizations within a region, community or hospital system. In practice the term HIE may also refer to the organization that facilitates the exchange. [10]

- **Patient Registries:** A patient registry is an organized. system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves a predetermined scientific, clinical, or policy purpose(s).[11]

- **Patient portal:** A patient portal is a secure online website that gives patients convenient 24-hour access to personal health information from anywhere with an Internet connection. Using a secure username and password, patients can view health information. [12]

- **Clinical Data warehouse:** A Clinical Data Repository (CDR) or Clinical Data Warehouse (CDW) is a real time database that consolidates data from a variety of clinical sources to present a unified view of a single patient. It is optimized to allow clinicians to retrieve data for a single patient rather than to identify a population of patients with common characteristics or to facilitate the management of a specific clinical department.

  Typical data types which are often found within a CDR include: clinical laboratory test results, Patient demographics, Pharmacy Information, Radiology reports and images, Pathology Reports, Hospital admissions, discharge and transfer dates, ICD-9 and ICD-10 codes and discharge summary and Progress Notes. [13]

- **Claim data from Payer: Claims data**, by its nature, is owned by the insurance company which covers each patient. It is nearly unheard of for a provider to have access to claims from all commercial insurance carriers used by all patients for whom they care.

- **Research studies:** A scientific study of nature that sometimes includes processes involved in health and disease. For example, clinical trials are research studies that involve people. These studies may be related to new ways to screen, prevent, diagnose, and treat disease. They may also study certain outcomes and certain groups of people by looking at data collected in the past or future. [14]

- **Genetic Database:** A genetic database is one or more sets of genetic data (genes, gene products, variants, phenotypes) stored together with software to enable user to retrieve genetic data, add genetic data, and extract information from the data. Genetic data base are repository of organized data that are a resource for understanding how organism function.[15]

- **Public record:** Any information, minutes, files, accounts or other records which a governmental body is required which a governmental body is required to maintain, and which must be accessible to security by the public. This include the files of most legal actions. A court will take "Judicial notice" of a public record introduce as evidence.[16]

- **Web search:** A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.[17]

- **Social media:** Social media are computer-mediated technologies that allow the creating and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks. The variety of stand-alone and built-in social media services currently available introduces the challenges of defining. [18]

- **Internet of things:** The Internet of Things (IoT) is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.[19]

- **Financial transaction:** A financial transaction is an agreement, communication, or movement carried out between a buyer and a seller to exchange an asset for payment. It involves a change in the status of the finances of two or more businesses or individuals.[20]

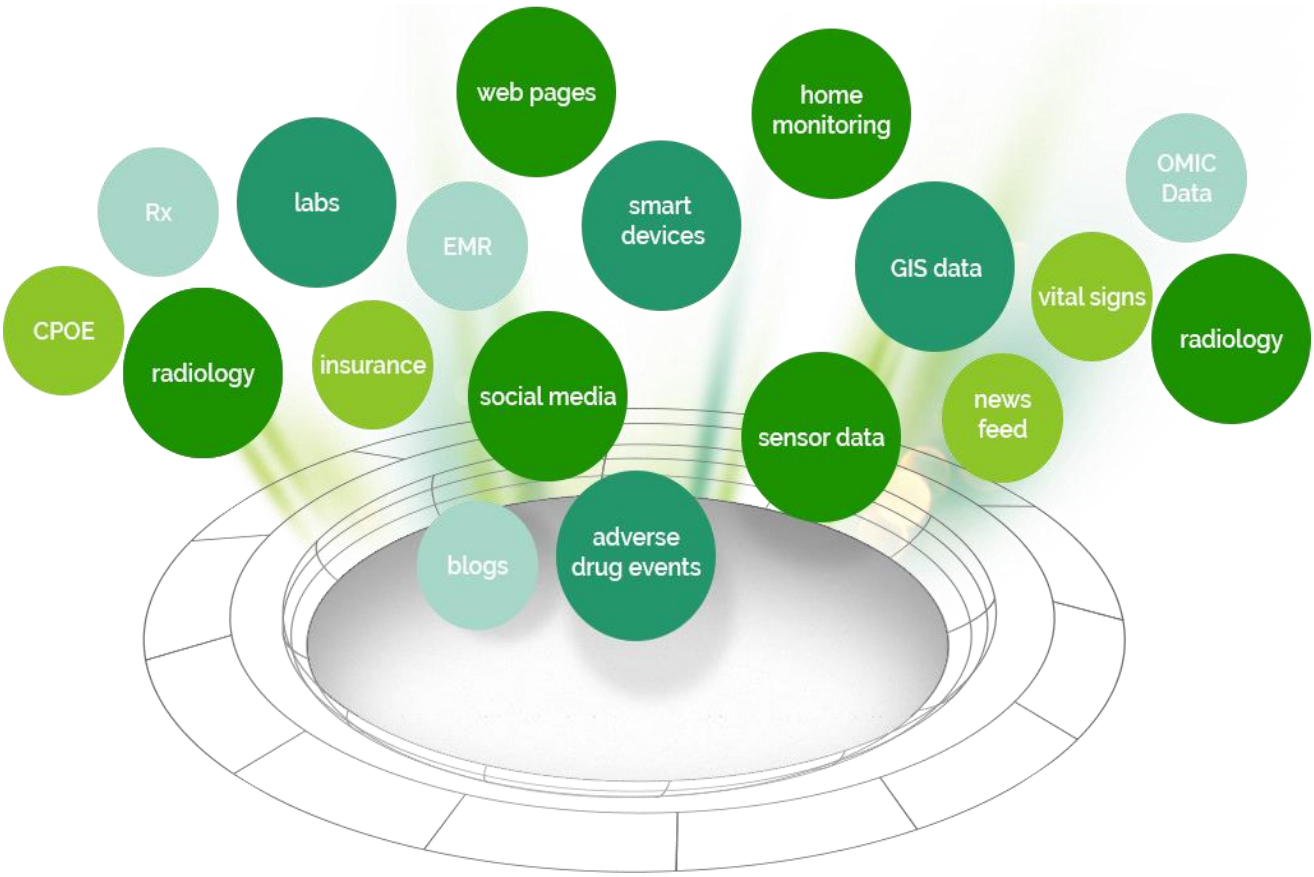


Fig. 4.1 Sources of Big Data

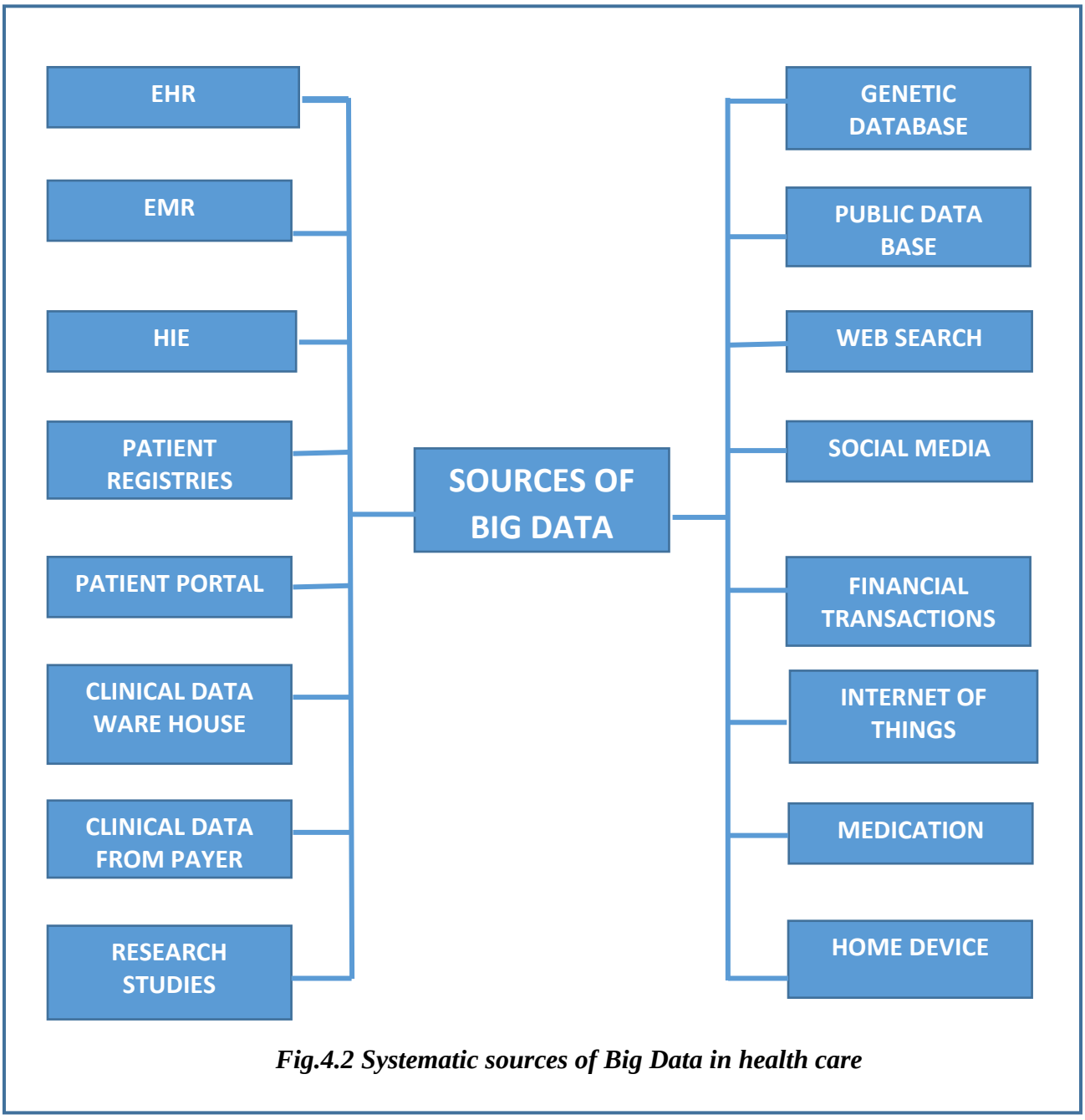


Fig.4.2 Systematic sources of Big Data in health care

4.5 Platform and Tools for Big Data Analytics in Healthcare:

- **The Hadoop distributed file system (HDFS):**

  The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project.[20]

  Hadoop is an open source implementation for large scale batch processing system. It uses the MapReduce framework introduced by google by leveraging the concepts of map and reduce functions as well know used in functional programming. Its framework is written in java to deploy custom- written programs and optimized for contiguous read requests, Where processing consist of scanning all data, response time can vary from minutes to hours, Hadoop with an advantage of its massive scalability, can process data fast. It leverage a cluster of nodes to run MapReduce program massively in parallel. MapReduce program consists of two steps:

  1. Map: To process input data

  2. Reduce: To assemble intermediate results into a final results.

Each cluster node has consist a local file system and local CPU to run the MapReduce programs. The local file system called Hadoop Distributed File System (HDFS).

Hadoop is being used for analysis of unstructured data such as log, text, and click stream, and email spam detection, index web searches, recommendation engine, prediction in financial services, genome manipulation.

- **Map Reduce**:

  Map Reduce has emerged as a popular way to harness the power of large clusters of computers. It allows programmers to think in a data-centric fashion: they focus on applying transformations to sets of data records, and allow the details of distributed execution, network communication and fault tolerance to be handled by the Map Reduce framework. And also Map Reduce is typically applied to large batch-oriented computations that are concerned primarily with time to job completion. The Google Map Reduce framework and open-source Hadoop system reinforce this usage model through a batch-processing implementation strategy: the entire output of each map and reduce task is materialized to a local file before it can be consumed by the next stage. Materialization allows for a simple and elegant checkpoint/restart fault tolerance mechanism that is critical in large deployments, which have a high probability of slowdowns or failures at worker nodes. We propose a modified MapReduce architecture in which intermediate data is pipelined between operators, while preserving the programming interfaces and fault tolerance models of previous Map Reduce frameworks. To validate this design, R. Abbott and H. Garcia Molina develop the Hadoop Online Prototype (HOP), a pipelining version of Hadoop. Pipelining provides several important advantages for Map Reduce framework; it also raises new design challenges.

- **Pig:**

  Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization .An application that creates map-reduce jobs based on a language called Pig Latin, which is workflow driven. It was originally created at Yahoo! (company).Apache Pig is good for structured data too, but its advantage is the ability to work with BAGs of data (all rows that are grouped on a key), it is simpler to implement things like:

  1. Get top N elements for each group;

  2. Calculate total per each group and then put that total against each row in the group;

  3. Use Bloom filters for JOIN optimizations;

  4. Multi query support (it is when PIG tries to minimize the number on MapReduce Jobs by doing more stuff in a single Job)

  5. Hive is better suited for ad-hoc queries, but its main advantage is that it has engine that stores and partitions data. But its tables can be read from Pig or Standard MapReduce.

  6. One more thing, Hive and Pig are not well suited to work with hierarchical data.

  Pig integration with streaming also makes it easy for researchers to take a Perl or Python script they have already debugged on a small data set and run it against a huge data set. PIG is best for semi structured data, for programming, used as procedural language, does not support partitions, don't have dedicated metadata of database. [21]

- **Hives:**

  Hives is data ware house software that facilitate querying and managing huge data residing in distributes storage. Instead of writing huge raw map reduce program in some programing language, hives provide a SQL like interface to data store in Hadoop file system. And there is another popular Hadoop eco system i.e. pig which is a scripting language with a focus on data flows hives provide data base query interface to apache Hadoop. People ask often why do pig and hives exist when they seems to do much of the same things, hives because of its SQL like query language is often used as interface to an apache Hadoop base data warehouse . Hives is considered friendlier and more familiar to user who are used to using SQL for querying data. Pig fits in though its data flows strength where it take on the task for bringing data into apache Hadoop and working it with to get it into the form for querying. A good overview how this works is in Alan gates posting on the yahoo developer blog titled pig and hive at yahoo, from a technical point of view both pig and hives are feature complete so you can do task in either tool. However you will find one tool or the other will be preferred by the different group that have to use apache Hadoop. The good part is they have choice and both tool work together. Since every network owner will depending on partner to get the service where they does not have the service tower. [22]

- **JAQL:**

  JAQL is a functional data query language, which is built upon JavaScript Object Notation Language (JSON). It is a data flow language that operates on unstructured, structured and semi structured data and is used for general purposes. It supplies a model for accessing data in traditional formats and provides functions for basic functions like filtering, aggregation and joins.

JAQL is extendable with operations written in many programming languages because JSON has a much lower impedance mismatch than XML for example, yet much richer data types than relational tables. Every JAQL program is run in JAQL shell. Initially JAQL shell is run by JAQL shell command. To achieve parallelism, it converts high-level queries into low-level queries consisting of MapReduce jobs. Similar to the flow of data typically present in a MapReduce job, JAQL queries can be thought of as pipelines of data that flow through various operators and end in a sink which is the final destination Ecosystem. The choice of a particular tool depends on the needs of the analysis, the skill set of the data analyst, and the trade-off between development time and execution time. Apache Pig provides a data flow language, Pig Latin that enables the user to specify reads, joins and other computations without the need to write a MapReduce program. Like Hive, Pig generates a sequence of MapReduce programs to implement the data analysis steps. JAQL is used more often for data processing and querying. [23]

- **Zookeeper:**

  In the history every application was a single program which runs on a single PC with an only CPU. Now days, situation is changed. In Cloud Computing & Big Data world, multiple applications are running independently on different set of computers in parallel with each other. So its tedious job for developers to maintain coordination among all these independent applications. It may cause failure at certain point of time. So to avoid this problem Zookeeper was designed. Zookeeper is a robust service. With the help of Zookeeper application developers can focus only on their application logic rather than coordination.

It exposes an easy API which helps developers to apply common coordination tasks like managing group membership, choosing a master server & managing metadata. Zookeeper is an application library. It has two principal API implementations- C and java. It has a service component made in java that runs on an assembly of dedicated servers. Due to assembly of servers, Zookeeper can increase throughput & tolerate faults or errors. Zookeeper is a building block for distributed systems. Zookeeper is a distributed coordination service. It is extremely consistent & highly available service. It is top level apache project created at yahoo. We can create distributed locks, distributed queues, group membership, master elections, distributed configuration & much more with the help of it. Various properties of zookeeper are operations are controlled, Updates are atomic & changes are robust. Zookeeper is an important part of HADOOP. Many HADOOP related projects like HBase, HDFS high availability & flume are based on it.

Zookeeper provides certain coordination services:

1. Name Service- This service is used to map a name to some data linked with that name. For example a telephone directory is a name service in which name of person is mapped to his or her telephone number. Another example is DNS service in which a domain name is mapped to an IP address. Zookeeper exposes easy interface to do that.

2. Locking- To make a serialized access to shared resources in distributed sys, we need distributed mutexes. With the help of zookeeper, we can implement it efficiently. Configuration management- Zookeeper centrally manage and store the configuration of your distributed system. This means that any new nodes joining will pick up the up-to-date centralized configuration from Zookeeper as soon as they join the system.

This also allows you to centrally change the state of your distributed system by changing the centralized configuration through one of the Zookeeper clients. [24]

- **HBase:**

  HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java much like a typical MapReduce application. HBase does support writing applications in Avro, REST, and Thrift. An HBase system comprises a set of tables. Each table contains rows and columns, much like a traditional database. Each table must have an element defined as a Primary Key, and all access attempts to HBase tables must use this Primary Key. An HBase column represents an attribute of an object; for example, if the table is storing diagnostic logs from servers in your environment, where each row might be a log record, a typical column in such a table would be the timestamp of when the log record was written, or perhaps the server name where the record originated. In fact, HBase allows for many attributes to be grouped together into what are known as column families, such that the elements of a column family are all stored together. This is different from a row-oriented relational database, where all the columns of a given row are stored together. With HBase you must predefine the table schema and specify the column families. However, it's very flexible in that new columns can be added to families at any time, making the schema flexible and therefore able to adapt to changing application requirements. [25]

- **Cassandra:**

  Cassandra is also a distributed database system. It is designated as a top-level project modeled to handle big data distributed across many utility servers. It also provides reliable service with no particular point of and it is a NoSQL system. [26]

- **Apache Oozie:**

  Apache Oozie is a scheduler system to run and manage Hadoop jobs in a distributed environment. It allows to combine multiple complex jobs to be run in a sequential order to achieve a bigger task. Within a sequence of task, two or more jobs can also be programmed to run parallel to each other. One of the main advantages of Oozie is that it is tightly integrated with Hadoop stack supporting various Hadoop jobs like Hive, Pig, Sqoop as well as system-specific jobs like Java and Shell. Oozie is an Open Source Java Web-Application available under Apache license 2.0. It is responsible for triggering the workflow actions, which in turn uses the Hadoop execution engine to actually execute the task. Hence, Oozie is able to leverage the existing Hadoop machinery for load balancing, fail-over, etc. Oozie detects completion of tasks through callback and polling. When Oozie starts a task, it provides a unique callback HTTP URL to the task, and notifies that URL when it is complete. If the task fails to invoke the callback URL, Oozie can poll the task for completion.

  Following three types of jobs are common in Oozie:

  • Oozie Workflow Jobs — these are represented as Directed Acyclic Graphs (DAGs) to specify a sequence of actions to be executed.

  • Oozie Coordinator Jobs — these consist of workflow jobs triggered by time and data availability.

  • Oozie Bundle — these can be referred to as a package of multiple coordinator and workflow jobs. [27]

- **Lucene:**

  Lucene is an open source java based search library. Lucene is very popular and fast search library used in java based application to add document search capability to any kind of application in a very simple and efficient way. [28]

- **Avro:**

  Avro is a remote procedure call and data serialization framework developed within Apache's Hadoop project. It uses JSON for defining data types and protocols, and serializes data in a compact binary format. Its primary use is in Apache Hadoop, where it can provide both a serialization format for persistent data, and a wire format for communication between Hadoop nodes, and from client programs to the Hadoop services.

  It is similar to Thrift and Protocol Buffers, but does not require running a code-generation program when a schema changes (unless desired for statically-typed languages).

  Apache Spark SQL can access Avro as a data source.

  An Avro Object Container File consists of:

  - A file header, followed by
  - One or more file data blocks.

  A file header consists of:

  - Four bytes, ASCII 'O', 'b', 'j', followed by 1.
  - File metadata, including the schema definition.
  - The 16-byte, randomly-generated sync marker for this file.

For data blocks Avro specifies two serialization encodings: binary and JSON. Most applications will use the binary encoding, as it is smaller and faster. For debugging and web-based applications, the JSON encoding may sometimes be appropriate. [29]

- **Mahout:**

Mahout is one of the more well-known tools for ML. It is known for having a wide selection of robust algorithms, but with inefficient runtimes due to the slow MapReduce engine. In April 2015, Mahout 0.9 was updated to 0.10.0, marking something of a shift in the project's goals. With this release, the focus is now on a math environment called Samsara, which includes linear algebra, statistical operations, and data structures. The goal of the Mahout-Samsara project is to help users build their own distributed algorithms, rather than simply a library of already-written implementations. They still offer a comprehensive suite of algorithms for MapReduce and many have been optimized for Spark as well. Integrations with H2O and Flink are currently in development. This version is very new, so there is no published literature on it at the time of this writing other than the initial announcement by the developer team introducing the new features. Because most of the old algorithm implementations are still included, the rest of this section focuses on versions 0.9 and earlier. Among the more commonly cited complaints about Mahout is that it is difficult to set up on an existing Hadoop cluster. Additionally, while a lot of documentation exists for Mahout, much of it is outdated and irrelevant to people using the current version. The lack of documentation, a problem common too many machine learning tools, is partially alleviated by an active user community willing and able to help with many issues. One problem with using Mahout in production is that development has moved very slowly; version 0.10.0 was released nearly seven and a half years after the project was initially introduced. The number of active committers is very low, with only a handful of developers making regular

commits. The algorithms included in Mahout Focus primarily on classification, clustering and collaborative filtering, and have been shown to scale well as the size of the data increases. Additional tools include topic modeling, dimensionality reduction, text vectorization, similarity measures, a math library, and more. One of Mahout's most commonly cited assets is its extensibility and many have achieved good results by building off of the baseline algorithms. However, in order to take advantage of this flexibility, strong proficiency in Java programming is required. Committer Ted Dunning noted "It's not a product. It's not a package. It's not a service. Batteries are not included." Some researchers have cited difficulty with configuration or with integrating it into an existing environment –. On the other hand, a number of companies have reported success using Mahout in production. Notable examples include Mendeley, LinkedIn, and Overstock.com, who all use its recommendation tools as part of their big data ecosystems. Overstock even replaced a commercial system with it, saving a significant amount of money in the process. [30]

4.6 Phases in the Big Data Analytics Process:

We can map steps taken up while performing BDA Process to the data mining knowledge discovery steps as follows:

1. Data Acquisition And Storage:

As already mentioned that knowledge (the info, the information) is fed to the system through several external sources like clinical data from clinical decision support system (CDSS), EHR, EMR, machine generated device knowledge, from wearable devices, national health register information, pharmaceutical corporation conducted drug connected information, social media Knowledge (info.) like twitter feeds, Facebook feeds, web page blogs, articles and lots of additional information, this knowledge or information is either keep in data ware house or data base. With advent of cloud computing, it's convenient to store such voluminous information or knowledge on the clouds instead of on physical disk. This can be addition price effective and manageable way to store knowledge (info.).

2. Data Cleaning:

The data that has been inheritable ought to be complete and should be in a structured format for effective analysis, typically its seen in healthcare data from flows like, several patient don't share their information or data fully, like knowledge or data regarding their dietary habits, weight and modus vivendi. In such cases the empty field ought to be handled bafflingly. Another example can be for e.g. for field like gender of person, there is at the most one amongst 2 values e.g. female or male , just in other value or no value is present then such entire ought to be marked and handle consequently. The information from sensor, medical images information, and social media information should be ought to be expressed during a structured kind for suitable analysis.

3. Data Integration:

The BDA process data accumulated across numerous platforms. The data will vary in metadata (the range of fields, type, and format). The complete data must be mass properly and systematically into a datasets which might be effectively used for data analysis purpose. This is often an awfully difficult task, considering the big volume and sort of variety of big data.

4. Data Querying, Analysis And Interpretation:

Once the data is integrated and cleansed, future steps is to questioned or query the data. A query may be straight forward or simple query for ex. What is a fertility rate for a particular region? Or may be complicated query like what percentage patients with polygenic disorder are likely to developed heart related issue in next five years?

Depending on the complexness of the query, the information analyst must select acceptable platform and analysis tools.

A large no. of open source and proprietary platforms and tools are available in market. Some of them are Hadoop, MapReduce, Storm, and Grid Grain. Big data databases like Cassandra, HBase, MongoDB, Couch DB, Orient DB, Territory, Hive etc.

5. Use Case Based On Big Data Analytics in Healthcare

5.1 Johns Hopkins use big data to narrow patient care:

At Johns Hopkins Medicine, big data and analytics are at the core of the organization's goal to tailor medical treatments and procedures to individual patients. Launched in 2012 as the Johns Hopkins Individualized Health Initiative, or Hopkins inHealth, the effort is a collaboration between Johns Hopkins University, which includes the medical school; the health system, and Johns Hopkins Applied Physics Laboratory. Johns Hopkins Medicine includes 1,192-bed Johns Hopkins Hospital, the system's flagship; five other hospitals, and 40 outpatient specialty and primary-care sites. The goal of Hopkins inHealth is to discover new scientific measurements and models to predict the trajectory of diseases in current patients as well as how each patient's unique genetic makeup is likely to respond to medical treatments and procedures. To fuel those discoveries, Johns Hopkins plans to mine data from myriad sources, including electronic health records, DNA sequences and digital images. "At its core, big data is about massive amounts of electronic patient information that can be mined to yield tailored medical results," explains Scott Zeger, director of Hopkins inHealth and a biostatistics professor at Hopkins Bloomberg School of Public Health. Based on those results, physicians—in collaboration with their patients—can develop high-quality and cost-effective treatment plans. To move from vision to reality, Hopkins in Health's leaders are bringing together the university's resources to maximize the potential of individual projects. Those resources include research funding of pilot projects; access to a range of hardware and software platforms on campus, including supercomputers, clinical cohort databases and measurement devices; and expertise in study design and data analysis.

Physicians—in collaboration with their patients—can develop high-quality and cost-effective treatment plans. To move from vision to reality, Hopkins in Health's leaders are bringing together

the university's resources to maximize the potential of individual projects. Those resources include research funding of pilot projects; access to a range of hardware and software platforms on campus, including supercomputers, clinical cohort databases and measurement devices; and expertise in study design and data analysis.

There are many other big data projects underway to improve healthcare outcomes in such areas as radiation oncology, autoimmune diseases, interventional cardiology, cancer screening, prostate cancer, and cystic fibrosis, among others. Although many of these projects are in the exploration stage, Oncospace, a SQL database and set of clinical support tools to improve treatment planning and medical outcomes in radiation oncology, is already being used in direct patient care. Oncospace includes electronic medical data on cancers of the head and neck, prostate, pancreas and lung for 2,300-plus patients. Data for each type of cancer is stored in a separate cohort database on the same server and using the same schema.

While the number of patients in the database is relatively small today, Zeger expects Oncospace to grow in size and its query-processing approach more complex as new types of data and "millions" of new patient cases are added not only from Johns Hopkins but other academic medical centers as well. The first predictive model Johns Hopkins developed in radiation oncology helps customize radiation treatment plans for cancer patients. "The goal of radiation therapy is to treat the cancerous tissues while sparing as much as possible the normal tissues that surround it," says Todd McNutt, director of clinical informatics for radiation oncology and molecular radiation sciences at Johns Hopkins, and the lead researcher and developer on the Oncospace project. In head-and-neck cancers, for example, physicians want to spare critical anatomy involved in swallowing and talking.

While the number of patients in the database is relatively small today, Zeger expects Oncospace to grow in size and its query-processing approach more complex as new types of data and "millions" of new patient cases are added not only from Johns Hopkins but other academic medical centers as well. The first predictive model Johns Hopkins developed in radiation oncology helps customize radiation treatment plans for cancer patients. "The goal of radiation therapy is to treat the cancerous tissues while sparing as much as possible the normal tissues that surround it," says Todd McNutt, director of clinical informatics for radiation oncology and molecular radiation sciences at Johns Hopkins, and the lead researcher and developer on the Oncospace project. In head-and-neck cancers, for example, physicians want to spare critical anatomy involved in swallowing and talking. [31]

5.2 Penn Health Sees Big Data as Life Saver:

The University of Pennsylvania Health System, like many large health organizations, has poured enormous resources into building an enterprise wide data infrastructure. With the foundation in place, the health system known as Penn Medicine is embarking on a big data project to expand its information horizons and develop predictive analytics to diagnose deadly illnesses before they occur.

Penn Medicine is a leading academic medical center. Based in Philadelphia, it consists of the Raymond and Ruth Perelman School of Medicine and the University of Pennsylvania Health System. The health system includes the Hospital of the University of Pennsylvania and Penn Presbyterian Medical Center, Chester County Hospital, Lancaster General Heal, Penn Wissahickon Hospice and Pennsylvania Hospital as well as a number of inpatient and other care services.

The open-source, big data initiative, called Penn Signals, is focused on building out an enterprise warehouse and enabling the data science team to create learning models from historical, at-rest data and then position those models into a real-time data stream, says Corey Chivers, a data scientist at Penn Medicine who is one of the leads on the project.

"Our goal is to build an infrastructure that can scale up to handle a huge variety of data sources within our system that contain information about the health of our patient population," Chivers says. "We started with the obvious candidates—our electronic health records and labs—to try to develop predictive models for severe sepsis and heat failure. But we have plans to increase the use of predictive models on the Penn Signals platform and make them available outside the organization."

The backbone is a homegrown enterprise data warehouse, called Penn Data Store. The warehouse contains data from clinical and administrative systems, including the three major clinical information systems at Penn Medicine: an outpatient EHR from Epic, which is used by 1,800 affiliated physicians; an inpatient EHR from Allscripts, which is used within Penn Medicine's five hospitals; and an enterprise laboratory information system from Cerner. In all, the warehouse stores over 4 billion rows of clinical data, with 2 million being added each day. For the big data effort, Chivers and the Penn Medicine's data science team used an ETL procedure to pull data from the enterprise warehouse into an open source database from MongoDB that provides flexibility for the machine-leaning applications the data science team utilizes. From there, observations made by the clinical staff are converted into time-series formats, or events, that can be analyzed by machine learning applications, Chivers says.

The team uses Python programming language, ZeroMQ messaging and the iPython Notebook computational environment to pull data sets and explore that data using dimensional reduction and machine learning. They then can save predictive models they've developed and ship them up into the real-time data stream as operational models, Chivers explains. For many health conditions, timing is everything. In the case of sepsis, every hour a patient goes undiagnosed increases the mortality rate by more than 7 percent, according to clinical studies, which also estimate that only 50 percent of septic shock patients receive effective therapy on time.

At Penn Medicine, the algorithm to detect when a patient was slipping into severe sepsis relied on analysis of six vital sign measurements and lab values with threshold rules. The Penn Signals predictive model takes into account more than 200 clinical variables. It has enabled Penn Medicine to detect 80 percent of severe sepsis cases within 30 hours of the typical onset of symptoms, Chivers says.

The heart failure predictive algorithm is enabling Penn Medicine to detect 20 percent more patients who are trending toward cardiac failure, and identifying a group of patients that is five times more likely to be readmitted after heart failure. Having that predictive information on hand means Penn Medicine clinicians can intervene earlier with at-risk groups and focus resources on those patients more likely to have ongoing heart issues. Communicating the output from the predictive algorithms is done via text messages that alert specific clinical staff when a patient's condition is heading in a dangerous direction. Penn Medicine also has developed a mobile app, called Caroline, which provides clinicians with a pared-down version of a patients' electronic health record containing clinical data related to the alert.

The data science team also uses the online visualization site Plotly, as well as visualization tools within the iPython Notebook environment, to provide clinic department heads and clinical floor

leaders with aggregate looks at the predictive data. “Our visualizations are under constant development because we want the clinicians to have a deep view of how these predictive models work and the information they utilize,” Chivers says. “We’ve worked closely with clinical staff to understand how we as data scientists can communicate with them.”

What's Next?

Penn Medicine now finds itself at a crossroad: it’s built an open-source framework that can handle the influx of health data and utilize predictive algorithms in real-time, but the volume, velocity and variety of data are ramping up quickly, Chivers says. “We’re planning to utilize new data streams—from wearable devices, telemetry devices and ICU monitors—and as we move toward that machine-generated data that’s coming in at much higher rate, we have to focus on scalability.” Penn Signals plans to take the infrastructure to that next level through an agreement with Intel to partner in the development of the company’s Trusted Analytics Platform, or TAP.

TAP is an open-source infrastructure built on a data layer that includes Apache Hadoop, Spark and other data components, as well as an analytics layer that includes a data science tool kit to simplify model development and an extensible framework to generate predictive approaches. Penn Medicine plans to deploy a 100-terabyte data stack via the TAP framework, Chivers says. The health system also plans to market Penn Signals to other health care organizations, he adds. “We want to get it out to other providers and find out what’s most valuable to them—is it something they’d like to deploy themselves, or would it be more useful as a platform for service? We don’t have an answer to that, but we built the platform using open-source tools so that it could be utilized beyond Penn Medicine [32]

5.3 Beth Israel Launches Big Data Effort to Improve ICU Care:

Like so many healthcare organizations, Beth Israel Deaconess Medical Center has massive amounts of data from its clinical care, education and research efforts, and the volume keeps growing. Putting all that data to good use, especially to improve care and patient outcomes, is a top priority for the Boston-based, 650-bed hospital, which treats hundreds of thousands of inpatients and outpatients each year.

This fall, the medical center, which is a teaching hospital of Harvard Medical School, will begin pushing live feeds of data into a custom application that caregivers can use to analyze risk levels in the intensive care unit (ICU) at any given time. Ultimately, the application will help Beth Israel Deaconess predict which patients are at risk of developing dangerous complications like infections, blood clots, or bleeds.

"We've been using big data for a while now. We've used clinical data to help predict, for example, hospital length of stay, and we have used demographics and other data to make strategic decisions such as hospital expansions," says Kenneth Sands, chief quality officer and a senior VP at Beth Israel Deaconess. "But much of the work around quality and patient safety has not been part of the big data exercise."

The project, known as Risky States, includes IT and care teams at Beth Israel Deaconess, scientists from the Massachusetts Institute of Technology (MIT) and human-factors experts at Aptima, which provides data collection, measurement, analytical, modeling and decision support systems. Work on Risky States, and its specific intensity index application, began more than two years ago, and required months of data extraction, translation and loading to prepare the data and to determine the appropriate data elements to include. It also involved the creation of a model to synthesize and correlate the data and the development of a user interface to render the data into actionable information that nurses and doctors could use to improve patient care.

Healthcare organizations are increasingly looking to leverage the large volume of data they are amassing. In its "Healthcare IT Vision 2015" report, Accenture identifies the "intelligent enterprise"—which makes use of big data—as one of five key trends in the healthcare industry. The report notes that 41 percent of health executives say the volume of data their organization manages has grown by more than 50 percent in the last year.

Risky States is just one component of Beth Israel Deaconess' overall initiative to eliminate preventable harm in critical care, an effort that is funded by a $5.3 million grant from the Gordon and Betty Moore Foundation. It includes Risky States and the intensity index application as well as two other applications: MyICU, a patient-centered portal for ICU caregivers and families of patients in the ICU, and Content-Sensitive Checklist, an application that guides caregivers through appropriate and routine patient care and that takes into account context and an individual patient's current state in the ICU. Aptima has been working with Beth Israel Deaconess on all three, according to Kevin Sullivan, VP of operations at Aptima.

The Data

Risky States relies on mostly structured data from ICU clinical information systems that continuously gather vital-sign data, such as blood pressure, electrocardiogram (ECG) and pulse oximetry physiologic metrics, says Pat Folcarelli, RN, PhD, who serves as senior director of patient safety at Beth Israel Deaconess. In addition, data comes from a custom-built hospital information system that holds patients' electronic medical records, labs and patient flow that provides data about the movement of patients to different departments such as radiology. Also, there is data from HR systems that are used for staffing, scheduling and payroll to provide information about which nurse was working when, how much training and experience the nurses

have and whether a nurse on the floor regularly works in a specific ICU or is working there as a substitute.

Accumulating the data hasn't been a problem. For many years, Beth Israel Deaconess has been using electronic applications including EHRs, networked medical devices and other systems that create a lot of data, and the hospital has about 3 petabytes of stored data, which continues to grow at a rate of 25 percent a year. The bigger challenge for Risky States, and other big data projects the hospital embarks on, has been normalizing the data and prepping it for analytics. The way clinical care is documented can vary greatly; for example hypertension, high blood pressure and elevated blood pressure are three different terms that describe the same condition.

"There has been a lot of data cleanup that needed to be done, and in the process, we've learned a lot about structured data, and quality of data," says Folcarelli. She says it took at least a year to normalize the data and determine the data points that would work well in the model. Statisticians and analysts worked with clinicians and nurses during this process.

The hospital's IT team uses scripts that extract data from the transactional systems—the HIS, the clinical ICU systems, the HR systems—on a regular basis. The extracts are sent to the hospital's clinical data warehouse, which is built with Microsoft SQL Server technologies. The extraction, transformation and loading (ETL) process is managed using SQL Server Integration Services (SSIS), a platform for building enterprise-level data integration and data transformations solutions. SSIS can be used to update data warehouses, clean and mine data, manage SQL Server objects and data and extract and transform data from a variety of sources -- such as XML data files, flat files, and relational data sources -- and then load the data into one or more destinations, according to Microsoft.

Beth Israel Deaconess has been using Microsoft SQL Server technologies for several years, and not too long ago it implemented data warehouses based on Microsoft SQL Server 2014 which features built-in data compression technologies to improve application performance, a decision that has cut query times and improved access to big data with a hybrid cloud solution and better business intelligence (BI) tools, such as Microsoft Azure HDInsight and Microsoft Power BI for Office 365. HDInsight is an Apache Hadoop implementation that runs in the cloud and Power BI provides a set of online analytics and reporting tools.

The Model

Beth Israel Deaconess worked with data scientists at MIT to create decision trees and develop the statistical data models for Risky States using retrospective data it had compiled on all ICU patients from 2012 to 2014. "The two years of data covered 1,800 patients and every 12 hour shift worked during that time. A 12-hour shift is defined as one unit, so for a patient who stays seven days there, that's 14 units, and there are data points associated with that. So there are hundreds and hundreds of data points per patient," explains Sands.

Risky States also has benefitted from the expertise of Harvard researchers who developed custom software tools and wrote open-source code to mine large clinical data sets. "Their work has helped our efforts, even if only indirectly," Sands says. "Some of the skill sets that allowed development of these software tools allowed us to put together the data set that we've used for the Risky States analysis."

The models will have to continually be updated, and going forward, that will be handled by Aptima. "All we've done so far is to show these are things that influence risk at a point in time," Sands adds.

Creating the models is a critical function of the project that helps the hospital better understand which particular sets of events—determined from analyzing a variety of data sets —are more likely to occur with particular harms, according to Aptima's Sullivan. "Perhaps there's an uncommon condition with certain patients that drives risk, or maybe having more junior-level staff than senior-level staff during a shift. Any combinations could cause risk to go up or down," he says.

With the models in place, Aptima designed a big data analytics application for Beth Israel Deaconess. The principal analytic technique used is a statistical approach known as recursive partitioning. "Basically, the way it works is by finding the factors that are most effective in splitting the data set up in ways that identify risk of harm," says Sands. The IT team uses Microsoft's SSIS to manage the process of sending data from the data warehouse to the analytics application built by Aptima.

The Caregiver Application

The analytics application Aptima has built features a user interface to present the risk levels and the factors that contribute to them so clinicians and nursing staff can do something about the risk at that time. Clinicians and the nursing staff in all seven of Beth Israel Deaconess' ICUs – medical, coronary, surgical, etc. – will have access to the application, which is web-based, to run risk reports on PCs and even handheld devices. The application is hosted at Beth Israel Deaconess, is built on a Java back-end (as are the MyICU and the Content-Sensitive Checklist applications), and works with the hospital's Microsoft SQL Server data warehouses.

Staff can query the application, and the application will calculate a risk score for the ICU or even a particular patient by comparing data from the patients' records and HR data with the data in the models. The scores are displayed on a visual dashboard to the ICU staff. Until now, the application has been using only dummy data against the models, but Beth Israel Deaconess is about to go live

with new data feeds pulled from the data warehouse into the analytical application (again, using SSIS) in real time, every 15 minutes. The live feeds will enable staff to make more proactive decisions, such as delaying a patient's surgery—decisions that could reduce risks.

"Over time, the hope is to reduce the prevalence of risks by making changes in care. Or, for example, by providing more training to junior staff, better managing staff workloads, or blending junior-level staff with more senior-level staff during certain times," Sullivan says.

Beth Israel Deaconess also expects the application will help it tap into care techniques that no one's thought about before by performing what-if analysis. "We only know what we know," says Folcarelli. "We'd like to build in capability that would let clinicians signal the app that perhaps the risk level and intensity is different than what the system is saying. Maybe risk is higher. Is the system missing something that we need to add in?" she adds.

The key is for big data, and the applications that leverage all that data, to augment the knowledge and expertise clinicians and nurses already have in patient care, not supplant it.

"This is the beginning," says Sands, "and it is all about patient care." [33]

6. Study Methodology

A scientific literature review search was conducted by using all the journals and articles available in order to know the impact of BIG Data analytics in Healthcare, and how's it improve the healthcare.

This chapter describe the methods used in attaining the selected objectives.

For the accomplishment of the study the following methods were used:

6.1 Scope of the work:

This study help to identify the Big Data analytics has been used for healthcare in the verge for improving healthcare. This current study analyses the potential benefits and challenges of Big Data analytics in healthcare, and their efficiency and efficacy in healthcare.

6.2 Study Settings:

- International Institute of Health Management and Research, New Delhi.
- NTT Data, Bengaluru.

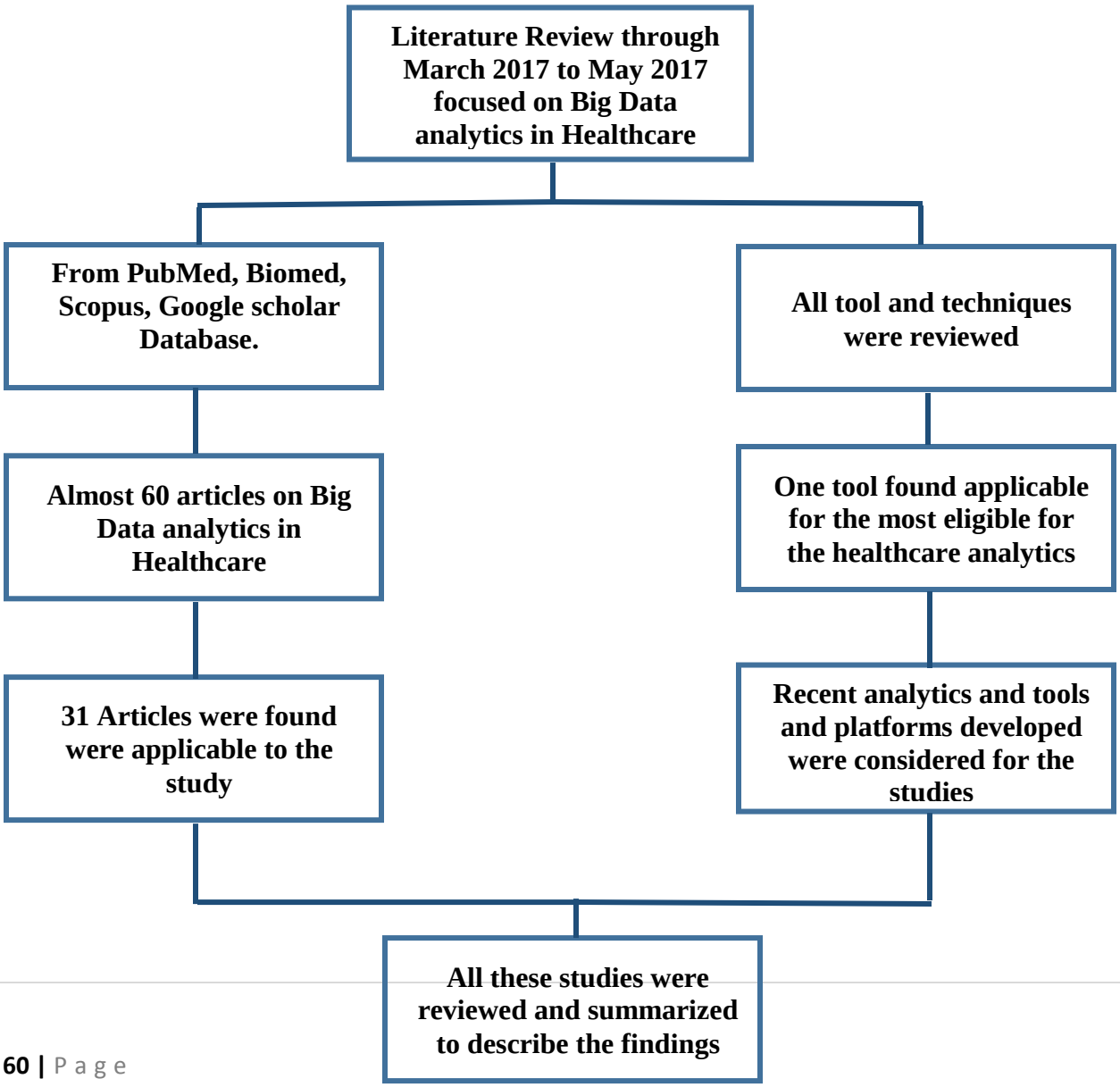
6.3 Data Sources:

This study is based on a systematic literature review of Big Data and Big Data Analytics and their relation to improvement the healthcare in patient health and other. Biomed, Google Scholar, Scopus and Wikipedia and Google Search were searched in March and April 2017, and PubMed and biomed and google for search engine for articles that discussed the Big Data analytics in Healthcare and Tools and techniques used in Healthcare. The Mesh Term used to indexed articles were: 1. "Big Data" 2. "Analytics in health care" 3. "Big Data Analytics" 4. "Tools and techniques for Big Data Analytics". The search terms only focus on the term synonymously. The search criteria did not include any limitation on publication date, earliest eligible article was published in past and 2017. The reference lists of included articles were also searched systemically.

6.4 Inclusion and Exclusion Criteria:

The study include literature on Big Data Analytics in Healthcare that improve healthcare. The inclusion criteria included Big Data, Big Data in healthcare. The tools and techniques are being used in healthcare. The articles from journals and also web pages were referred for the study purpose. The exclusion articles are those which are not in English or are in any other languages. The old manual methods for analytics, some basic health techniques in daily activities.

6.5 Study selection and Data Extraction:



7. Results:

7.1 Potential of Big data in healthcare:

Big data analytics has the potential to remodel the means of life sciences and health organization use subtle technologies to achieve insight from their clinical and alternative data repositories to form conversant selections. Analytics enable organization to analyze and explore data to spot relationships, trends and patterns, to reveal insight that. Once combine with the business context, create information, within the future, the implementation and usage of massive data can unfold speedily.

As big data analytics become additional thought, issues like guaranteeing privacy, safeguarding security, establishing standards and governance, and regularly improving the tools and technologies got to be resolved in a compliant and cost efficient manner only then organizations garner the advantages from big analytics in health and life science.

Big data additionally has the potential to considerably alter the business of health care delivery. Healthcare is heavily regulated. Health care provider needed are needed to produce an extraordinary level of transparency for justification for their activities. They can leverage big data technology not solely to boost services but also to improve operational efficiency and increase profitability.

Big data analytics has the potential to improve care, save lives, lower cost.

Healthcare and Payers:

- Analyzing patient characteristics and also the value and outcomes of care to spot the fore most clinically effective and cost effective diagnoses and treatment.
- Characterizing, identifying, predicting and minimizing fraud by implementing frauds detection and checking the accuracy and consistency of claims.

- Analyzing numbers of claims request quickly within the pre adjunction part to reduce fraud, waste and abuse.

Evidence-Based Medicine:

- Combining and analyzing a spread of structured and unstructured data-EMRs, clinical data, and genomic data, operational data- to match outcome, predict patient at a risk of disease or readmission, and provide a most effective care to reduce cost**.**
- Applying advance analytics to patient profile (e.g. segmentation and prognostic modeling (predictive modeling). To identify individual who would benefit from proactive care or lifestyle changes. For ex. Those patients at risk of developing a specific disease who would benefit from preventing care life style changes.
- Using historical data to individualize medical care by predicting and or estimating development or outcomes, such types of patient can select or choose elective surgery, will not benefit from surgery, are at the risk of medical complications or hospital acquired sickness or will have potential to co morbid conditions.
- Executing the gene sequencing are most effective to make genomic analysis a part of regular medical care decision process and its growing medical care.

Real time Healthcare and Clinical Analytics:

- Collection and publication of data on innovative medical treatment and procedures, it helping patient in deciding the care protocols or regimens that supply the most effective worth.

- Synthesizing and aggregating patient clinical records and claims data sets in real time, to provide services and related data to third parties. For ex. To identify patient inclusion in clinical trials, to provide licensing data to assist pharmaceutical companies.
- Sleuthing individual and population trends accurately and rapidly by deploying mobile applications that facilitate patients to manage their care, to find provider, and improve their health.
- Monitoring an observance medical device, as well as wearable, to capture and analyze in real time fast moving and large amount of data, for safety observance and adverse event prediction, enabling payers to watch adherence to drug and treatment regimens and observe trends that cause individual and population wellbeing edges.

Research and Development:

- Lower the attrition by improving predictive molding and produce a liner faster and more targeted research and development pipeline in medicine and devices.
- Investing applied statistical and mathematical tools and algorithms to enhance clinical trial design and patient requirement to higher tailor treatment to individual patient, thus reducing trail failures, and speeding new treatment to market.
- Analyzing clinical trials and patient records to spot follow up indication and check adverse effect before product reach to the market.

Public Health:

- Analyzing disease pattern and trailing diseases outbreaks and transmission to enhance public health surveillance and speed response.

- Rising data model to predict virus evaluation. Resulting a lot of accurately targeted vaccines. Turing large amount of data actionable information that can be used to identify needs, predict and stop crises, particularly for the advantage of populations.

7.2 Benefits:

Big data analytics has offered a brand new way to develop actionable insight, organize their future vision, maximize the outcome and scale back time to worth. This approach is additionally useful to predict perceptive info to the healthcare enterprise regarding their management, planning and the measurement. The evaluate result can enhance or help to enhance the decision making capacity of the top management.

- By recording disease outbreak and analyzing disease pattern, public health problem and issues may be improved with analytics approach. Large amount of data can facilitate to confirm desire and predict and forestall the long run or future crises.
- The EMR contain the standards (unstructured and structured) health information data that may be evaluated with the data analytic approach to predict patient condition that patient is in risk provide him/her effective care.
- Advance analytics can be applied to patient profile for identify individual, who can get benefits from predictive approach. Thus this may change lifestyle.
- Big Data analytics approach can be effectively enclosed in genomics analytics. To make this approach a section or a care of medical decision process.
- This data analytics approach can also help in analyze larger range of claim request to curtail down fraud cases. A good analysis will facilitate scale back fraud, waste and abuse.

- It can also be used to analyze real time huge/large volume of brisk data in hospitals. This approach could facilitate in the hospital to safety monitoring and negative event prediction.
- Big data technology permits facility manager to try side by side examination of performance report in area, staring from power monitoring to security services. It goes without saying that as facility performance goes up, so will patient safety, patient satisfaction, and facility rating.
- This is often another case wherever addition and more information neutrally ends in higher performance. Enhance observational and monitoring capabilities as a part of a big data solution make it thus healthcare facility manager and staff can pin point the area of with high energy use, and confirm where it is extremely needed and wherever energy is being wasted, and so take step to combat those inefficiencies. A decrease in overall energy use come with the advantage with slashed energy cost. It also likely to create more environmentally.
- With internet and cloud storage technologies perpetually improving and expending big data and can be accesses from anyplace. This is a great benefit and to doctor can access electronic medical records from anywhere at any time to enhance patient care.
- Big data is even ever changing however patient select their care provider or healthcare provider.
- Access live updates alerts permits healthcare provider within the facility to reply quickly and expeditiously to problem and concerns. With digital hospital solutions that send the correct data and correct information to the right individual in real time.so that user can make informed decision and get ahead to the issue.

- Big data analytics can help to cut down on administrative costs: as super computer we will use so maximum things we can do with the help of computers so the need of manpower will decrease so administrative cost will cut down.
- Big data analytics can use for clinical decision support system.
- Big data analytics improve patient experience as he/gets best healthcare service so this tends him/her to a level of good quality and satisfaction.
- Improve overall public health by capturing real time healthcare data that help the government to provide right cation of care at any outbreak.
- Medical error and wrongful death are the serious problems of healthcare. Misdiagnosis is main cause in the healthcare. There are many causes that could lead to a misdiagnosis but some truly are preventable, such as those relating to the misinterpretation of patient conditions, and over sights of medical history, these types of unaware mistakes made in healthcare. This is where analytics come in scenario, analytical tool integrate source from multiple sources, and also interpret the seemingly unrelated data and find previously unrelated relationships. Through analytical tool misdiagnosis can essentially be eliminated.

7.3 Healthcare Big Data Analytics Challenges:

- **The deluge of digital data:** The amount of data is generating. This amount of data is almost double in less than decade, the sheer size of this data is a major challenge.
- **Not enough men Power related to big data analytics:** The McKinsey Global Institute estimates that there will be a 100,000-plus-person analytic talent shortage at least through 2020, which could mean 50–60% of data scientist positions may go unfilled. Data scientists need more than the highly technical skillsets held by today's data analysts.

They must have well-developed soft skills such as communication, collaboration, leadership, creativity, and more. Of the healthcare leaders surveyed in July 2014, 60% were unsure whether they had in-house expertise necessary [34]

- **Data and Information Privacy challenge:** Privacy problem became progressively imperative recently. As web transactions and combinations. Cloud storage, mobile devises expose maximum personal data to potential misuse. Whereas online and social media user are rather inconsistent about privacy implication of their own behavior.
- The major issue is the selection of suitable implementation platform. It should support, at a minimum, the key function necessary for processing the data.
- Other challenge of big data analytics in big data that there are lots of data analytics tool in market. Which tool is right and can help in gaining benefits. Selection of tools in big data analytics is another challenge in big data analytics.
- Scalability and dynamics are the challenges in big data analytics.
- Unstructured clinical notes in healthcare, understanding of these clinical notes in right context is another challenge.
- Capturing the several data through several sensor.
- **Sharing patient Information/Data:** Centralizing patient record is defiantly effective within a healthcare workplace or hospital. But there is concern about sharing information or data with outside practitioners. Effectively sharing complete medical record and integration of complete different medical information/data is a challenge. Less integration of these medical record is a challenge in analytics.
- Ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance is another challenge.

- Real time data collection is a key of big data analytics in healthcare. The leg between collections a processing of data is need to be addressed.
- Governance and standards are need to be consider.

8. Conclusion

Big Data analytics has the potential to remodel the healthcare system use to refined technologies to gain insight to their clinical and alternative data repositories and create to make good decision. In the future, wide spread implementation and uses of data analytics across the hospital and healthcare field helps to deal with the healthcare issues and challenges.

Big Data include analytics tools and predictive analytics tools that helps health care system to reporting to predicting at earlier stages, I have discussed in this paper, the vison of big data that how big data collecting analyzing and managing healthcare data if different forms from multiple sources.

Big data analytics is also helping the Healthcare payer, evidence based monitoring, real time data collection and monitoring, clinical analytics and research and development, public health that can change the whole scenario of healthcare. Which define the new path way:

Right living:

Patient will build worth by taking an active role in their own treatment, as well as diseases prevention. The right living path way focus on the encouraging patient to from better decisions that facilitate them to stay healthy. Like correct diet and exercise, and take an active and energetic role in their own care if they become sick.

Right care:

Pathway involves guaranteeing that patient get the foremost timely, applicable treatment available, right care needed a coordinate approach: across settings and healthcare providers. All care giver ought to have the same information and work towards the same goal to duplication of effort and suboptimal strategies.

Right Provider:

This pathway proposed that patient must always be treated by high performing professionals that area unit best match to the task and can reach to the most effective outcome.

Right provider has two meanings:

- Right match of provider skill set to the complexity of assignment, e.g. nurse or physician assistant activity task that don't need a doctor.
- Specific choice of the provider with the most effective outcomes.

Right Value:

To meet the goals of this pathway, providers and payer will continuously enhance healthcare value while preserving or improving its quality. This pathway might involve multiple measure for granting cost effectiveness of care. Such as tying provider reimbursement to patient outcomes, or eliminating fraud, waste, or abuse in the system.

Right Innovation:

This pathway involves the identification of latest therapies and approaches to delivering care, across all aspects of the system, and rising the innovation engine themselves. To capture this worth, stakeholders should create higher use of previous trial data, such as by looking for high potential targets and molecules in pharma. The data can be used to opportunities to boost clinical trials and traditional treatment protocols, as well as both for birth and inpatient surgeries.

9. Recommendation

- Governmental interference: governmental interference is important if government will take interest in big data analytics in this situation government should set standards and protocol related to big data and analytics so we can overcome the challenges of big data.
- Government can give license to selected companies if they are meeting with the standards related to big data and analytics.
- By opening no. of institute and colleges related to big data and analytics we can overcome the challenges part.
- By promoting new talent and data storage techniques.
- By making standards related to big data and analytics we can solve many challenges related to big data analytics.
- Promotion of Big Data analytics and adoption and benefits of Big Data analytics.

These are few suggestion trough which we can overcome the challenges of big data analytics and we get advantages of big data analytics.

References:

1. Big data: Bijesh Dhyani, Anurag Barthwal, Big Data Analytics using Hadoop, International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014.

2. J.Ramsingh, Dr.V.Bhuvaneswar, an Insight on Big Data Analytics Using Pig,Scrip0074, Volume 4, Issue 6, November - December 2015, ISSN 2278-6856.

3. Bonnie Feldman, Ellen M. Martin, Tobi Skotnes, October 2012, Big Data in Healthcare Hype and Hope.

4. Bonnie Feldman, Ellen M. Martin, Tobi Skotnes, October 2012, Big Data in Healthcare Hype and Hope.

5. S. G. Nandhini, V. Lavanya, K.Vasantha Kokilam, Big Data Analytics in Health Care, September 2015 | IJIRT | Volume 2 Issue 4 | ISSN: 2349-6002.

6. Wullianallur Raghupathi, and Viju Raghupathi, Big data analytics in healthcare: promise and Potential, 2014, 2:3http://www.hissjournal.com/content/2/1/3.

7. Mu-Hsing Kuo, Health big data analytics: current perspectives, challenges and potential solutions. Vol. 1, Nos. 1/2, 2014.

8. Electronic health Recordswww.cms.gov/Medicare/EHealth/EHealth/EHealthRecords.

9. Electronic Medical Records, http://whatis.techtarget.com/definition/electronic-medical-record-EMR

10. Health Information Exchange, https://en.wikipedia.org/wiki/Health_information_exchange

11. Patient Registries, www.pcori.org/assets/11-Gliklich-Slides-Registries.pdf

12. Patient Portal,https://www.healthit.gov/providers-professionals/faqs/what-patient-portal

13. Clinical Data Repository, https://en.wikipedia.org/wiki/Clinical_data_repository

14. Research, https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=651211

15. Genetic Database, http://www.nature.com/subjects/genetic-databases

16. Public Record, http://legal-dictionary.thefreedictionary.com/Public+records

17. Web Search Engine, https://en.wikipedia.org/wiki/Web_search_engine

18. Social Media, https://en.wikipedia.org/wiki/Social_media

19. **Internet of Things**,http://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT

20.**FinancialTransactions,https**://books.google.com/books?id=eLXFqy8vAgC&pg=PT231&lpg=PT231&dq

21. Ahmed Eldawy, Mohamed F. Mokbel "A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data "Proceedings of the VLDB Endowment, Vol. 6, No. 12 Copyright 2013 VLDB Endowment 215080

22. N.Pushplata , P.Sudheer " Data processing in big data by using hives interfaces, volume 3, issue 4,April 2015.

23. Misha Shetha, Purna Mehtab , Khushali Deulkar "Enhancing Massive Data Analytics with the Hadoop Ecosystem" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3, Issue 11 November, 2014 Page No. 9061-9065

24. Smita Konda, Rohini More "Balancing & Coordination of Big Data in HDFS with Zookeeper and Flume" Volume: 02 Issue: 09 | Dec-2015, e-ISSN: 2395-0056, p-ISSN: 2395-0072

25. Shalini Sharma, Satyajit Padhy, Improvising Data Locality and Availability in Hbase Ecosystem IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. XI (Mar-Apr. 2014), PP 36-41

26. Wullianallur Raghupathi, and Viju Raghupathi, Big data analytics in healthcare: promise and Potential, 2014, 2:3http://www.hissjournal.com/content/2/1/3

27. Apache Oozie, www.tutorialspoint.com/ Apache Oozie

28. lucence,www.tutorialspoint.com/ lucence

29. Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter* and Tawfiq Hasanin "A survey of open source tools for machine learning with big data in the Hadoop ecosystem" Landset et al. Journal of Big Data (2015) 2:24 DOI 10.1186/s40537-015-0032-1

31. Linda Wilson, February 04 2016, 5:45am EST, Johns Hopkins use big data to narrow patient care. https://www.healthdatamanagement.com/news/johns-hopkins-uses-big-data-to-narrow-care

32. Greg Gillespie, Published January 11 2016, 7:22am EST, Penn Health Sees Big Data as Life Saver.

33. Beth Bacheldor, Published September 16 2015, 7:55am EDT, Beth Israel Launches Big Data Effort to Improve ICU Care.

34. http://www.mckesson.com/healthcare-analytics/healthcare-big-data-challenges/#footNot

Other reading articles:

[1] Manish Kataria1, Ms. Pooja Mittal, Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql, ISSN 2320–088X, IJCSMC, Vol. 3, Issue. 7, July 2014, pg.759 – 765.

[2] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT, BUSINESS INTELLIGENCE RESEARCH.

[3] Christopher Olston, Benjamin Reed Utkarsh Srivastava, Ravi Kumar Andrew Tomkins, Pig Latin: A Not-So-Foreign Language for Data Processing.

[4] Bijesh Dhyani, Anurag Barthwal, Big Data Analytics using Hadoop, (0975 – 8887) Volume 108 – No 12, December 2014.

[5] Sneha Mehta, Viral Mehta, Hadoop Ecosystem: An Introduction, ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

[6] Harshit Kumar, Nishant Singh, Review paper on Big Data in healthcare informatics, Volume: 04 Issue: 02 | Feb -2017, e-ISSN: 2395 -0056, p-ISSN: 2395-0072.

[7] S. G. Nandhini1, V. Lavanya2, K.Vasantha Kokilam, Big Data Analytics in Health Care, September 2015 | IJIRT | Volume 2 Issue 4 | ISSN: 2349-6002

[8] Anand Loganathan, Ankur Sinha, Muthuramakrishnan V,and Srikanth Natarajan, A Systematic Approach to Big Data Exploration of the Hadoop Framework, ISSN 0974-2239 Volume 4, Number 9 (2014), pp. 869-878

[9] J.Ramsingh, Dr.V.Bhuvaneswari, An Insight on Big Data Analytics Using Pig Script, Volume 4, Issue 6, November - December 2015 ISSN 2278-6856

[10] Bhawana Sahare, Ankit Naik2, Kavita Patel, Study of HADOOP, Volume 2 Issue 6, Nov-Dec 2014.

[11] Shalini Sharma, Satyajit Padhy, Improvising Data Locality and Availability in Hbase Ecosystem, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. XI (Mar-Apr. 2014), PP 36-41.

[12] Sumaiya Nazneen1, Sara Siraj2, Tabassum Sultana3, Nabeelah Azam4, Tayyiba Ambareen5, Ethemaad Uddin Ahmed6, Mohammed Salman Irshad7, Influence of Hadoop in Big Data Analysis and Its Aspects

[13] Lidong Wang1, Guanghui Wang, Cheryl Ann Alexander, Big Data and Visualization: Methods, Challenges and Technology Progress, *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.

[14] Priyanka K, Prof Nagarathna Kulennavar, a Survey on Big Data Analytics in Health Care, ISSN: 0975-9646.

[15] Rebecca Hermon, Patricia A H Williams, Big data in healthcare: What is it used for? 2014, DOI: 10.4225/75/57982b9431b48.

[16] Wullianallur Raghupathi, Viju Raghupathi, Big data analytics in healthcare: promise and Potential, doi: 10.1186/2047-2501-2-3, 014, 2:3http://www.hissjournal.com/content/2/1/3.

[17] Sanjeev Dhawan , Sanjay Rathee , Big Data Analytics using Hadoop Components like Pig and Hive, ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629.