Big Data in Healthcare Management: A Review of Literature

Senthilkumar SA1, Bharatendara K Rai2, Amruta A Meshram2, Angappa Gunasekaran3, Chandrakumarmangalam S4

1Department of Management, Pondicherry University, Pondicherry, India
2Charlton College of Business, University of Massachusetts Dartmouth, North Dartmouth, USA
3School of Business and Public Administration, California State University, Bakersfield, USA
4Department of Management Studies, Anna University Regional Campus, Coimbatore, India

Email address:
senthilkumarSa@yahoo.co.in (Senthilkumar SA)

To cite this article:
doi: 10.11648/j.ajtab.20180402.14

Received: March 4, 2018; Accepted: May 3, 2018; Published: July 3, 2018

Abstract: A systematic literature review of papers on big data in healthcare published between 2010 and 2015 was conducted. This paper reviews the definition, process, and use of big data in healthcare management. Unstructured data are growing very faster than semi-structured and structured data. 90 percentages of the big data are in a form of unstructured data, major steps of big data management in healthcare industry are data acquisition, storage of data, managing the data, analysis on data and data visualization. Recent researches targets on big data visualization tools. In this paper the authors analysed the effective tools used for visualization of big data and suggesting new visualization tools to manage the big data in healthcare industry. This article will be helpful to understand the processes and use of big data in healthcare management.

Keywords: Big Data, Data Acquisition, Data Storage, Data Analytics, Data Visualization, Healthcare Management

1. Introduction

Healthcare Industry is one of the world’s biggest and widest developing industries. During, the recent years the healthcare management around the world is changing from disease-centered to a patient-centered model [1] and volume-based to a value-based healthcare delivery model [2]. Educating the superiority of health care and decreasing the cost is a principle behind the developing movement toward value based healthcare delivery model and patient-centered care. The volume and demand for big data in healthcare organizations are growing little by little [3]. To provide effective patient-centered care, it is essential to manage and analyze huge health data. The outdated data management implements are not sufficient enough to analyze big data as variety and volume of data sources have increased in the past two decades. There is a need for new and innovative big data tools and technologies that can meet and exceed the ability of managing healthcare data [4]. Research study predictions on the worldwide big data expenditure in the healthcare business to progress towards Compound Annual Growth Rate (CAGR) of 42% during this years 2014-2019 [5].

The big data are used to predict the diseases before they emerge based on the medical records. Many countries’ public health systems are now providing electronic patient records with advanced medical imaging media [6]. The practice of big data takes the prospective to encounter the upcoming market needs and trends in healthcare establishments [7]. Big data provides a great opportunity for epidemiologists, physicians, and health policy experts to make data-driven judgments that will eventually develops the patient care [8]. The authors have used Google trends for analyzing the ‘big data in healthcare’ between 2010 and 2015. The resulting graph is shown in figure 1.
Google trends is a unique, liberally available online portal of Google Inc. that permits consumers to cooperate with Internet hunt data, which may provide profound visions into peoples activities and health-connected occurrences. Google flu trends and Google dengue trends are used for nowcasting the spread of diseases like flu and dengue. Google trends have been used in several research publications [9].

It can be seen from Figure 1 that the term ‘big data in healthcare’ really took off around early 2013. Increase in interest in this term can be related to a popular report by McKinsey & Company that came out in early 2013 [10]. The report highlights that healthcare expenses contributes about 17.6% of GDP and have a potential to reduce healthcare spending by $300 billion to $450 billion.

Currently, most big data studies in healthcare merely concentrate on technological understanding of big data [11]. Only a few topics had been debated the big data analytics in healthcare Information Technology [12]. The main concept of this is to review the processes and applications of big data in healthcare management.

This content is structured as follows: Section-2 introduction to the definitions and different dimensions of big data in healthcare. Section-3 discusses the search strategy and the steps in literature review of articles dealing with ‘big data in healthcare’. In section-4, the process of big data management in healthcare is presented. Section-5 discusses the applications, benefits, and challenges of big data in healthcare. Since the privacy issues have been increasing recently, the laws related to protection and management of health and medical data are also discussed in this section. Finally the content finishes with conclusions and recommendations for upcoming investigation in the section-6.

2. Definition of Big Data

There are many definitions that have been provided by researchers for big data [13-14], however no sole definition of big data is generally known [15]. Baro et al [16] have done extensive research on the definition of bigdata and have proposed that a dataset could be qualified as “big dataset” only if Log (n*p) is superior or equal to 7. Report submitted by U. S. Congress in August 2012 explains big data as “large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information” [17]. In this content, we propose a new explanation of big data in healthcare.

The big data in healthcare involves collecting large collections of data from various healthcare foundations followed by storing, managing, analyzing, visualizing, and delivering information for effective decision making.

The big data in healthcare is associated with six characteristics viz., volume, variety, velocity, veracity, variability, and value. Various researchers have discussed first three V’s (volume, variety, and velocity) that are widely used for characterizing big data in their papers [18-19]. The remaining three V’s (veracity, variability, and value) are also discussed extensively by several researchers [20-21]. These six V’s of big data are depicted visually in Figure 2.
Volume denotes to the large quantity of data produced by the organization. Today healthcare data are in terabytes (10^{12} bytes), petabytes (10^{15} bytes) or Exabyte’s (10^{18} bytes)” [22]. In future the vast entirety of clinical data records will increase to zettabyte (10^{21} bytes) or yotta byte (10^{24} bytes). Such huge amount of data creates storage and massive analysis issue.

Variety arises from broadly contrasting bases of data or mash-ups of data resultant from autonomous sources in format [23]. The different format of data in healthcare can be classified as organized, semi-organized, or unorganized data [24]. Organized data includes laboratory data, clinical, sensor data and data from relational databases [25], semi-organized data includes data that is stored in Extensible Markup Language (XML) format [26], and unorganized data are free text data that usually does not have a precise design such as manual written note [27-28], data from X-ray images, radiological images and other medical imaging [29-30], Electronic Medical Record (EMR/HER) [31], graphics, patient discharge summaries, physiological measures (signals), healthcare data from social media and mobile phones [32-33]. 90 percent of big data are in the form of an unstructured data [34].

Velocity refers to the massive frequency during the current data is created, supplied and managed [35]. Velocity thus includes both equivalent the rapidity of data manufacture and the rapidity of data handling to meet demand. Accelerated increase of data is the third characteristic of big data [28, 29, 36]. The data that are generated can be either batch or real-time data [28]. The data’s contents are frequently fluctuating through the concentration of corresponding data assemblies, the summary of previous data or inheritance gatherings, and the different forms of streamed data from multiple sources [25]. Example of velocity can be ageing of the population that is constantly increasing giving rise to number of patients, which increases growth rate of data by 55–60% every year [30].

Variability refers to data fluctuations throughout the handling and lifecycle. Developing range and variability also grows the attraction of data and the possibility in providing valuable information, unforeseen, and hidden [20].

Value is the method of extracting valuable information from huge sets of data and it is usually referred to as big data analytics [25]. Data value is useful for proper making decisions [39-21].

McKinsey & Company [10] believe that transforming of data deals with what is suitable or right healthcare environment and right for the for a patient They have considered following five strategic ways to significance, based on the concept that value is resulting from the balance of patient impact (outcomes) and healthcare spend (cost).

Right living: Patients must be stimulated to be encouraged by taking an intimation part in their own health.

Right care: Patients must take the utmost timely, suitable treatment offered.

Right provider: Any professionals must have strong performance records who treats the patients and be skilled enough for succeeding the best results.

Right value: Suppliers and customers should frequently look for ways to expand value while maintaining or improving health-care quality.

Right innovation: Investors must concentrate on classifying new therapies and approaches to health-care delivery.

3. Methods

A complete literature analysis of article deals with ‘big data in healthcare’ related study were directed. A detailed search of publication of papers about big data in healthcare between 2010 and 2015 was used for review. The Table 1 includes the peer-reviewed journal articles from the large publishers’ viz. Science Direct, PubMed, Springer, Taylor, & Francis, Inderscience and other reports on big data in healthcare.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Science Direct</td>
<td>1,323</td>
<td>1,615</td>
<td>1,772</td>
<td>2,262</td>
<td>2,645</td>
<td>9,617</td>
</tr>
<tr>
<td>PubMed</td>
<td>65</td>
<td>72</td>
<td>127</td>
<td>190</td>
<td>210</td>
<td>664</td>
</tr>
<tr>
<td>Springer</td>
<td>7</td>
<td>17</td>
<td>28</td>
<td>40</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Taylor &amp; Francis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>42</td>
<td>51</td>
</tr>
<tr>
<td>Inderscience</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Others</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Grant total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10496</td>
</tr>
</tbody>
</table>

The detail search result yielded 10,496 papers. Our strategy for selecting the final review papers is explained in Figure 3. After reviewing these articles we discarded most of these papers.
After removing irrelevant entries and duplications, 573 papers were included for title review. After reading the title, 459 papers on big data that do not use the word healthcare in their title were excluded. Abstract reviewed by a human for eligibility, 36 papers were excluded since they were not directly related to healthcare and 78 papers relevant to healthcare full text criteria were reviewed. Finally, the remaining 76 papers met all the inclusion criteria were retained. The Figure 3 shows the steps followed in search strategy.

4. Process of Big Data Analysis in Healthcare Industry

Big data analysis has the prospective to change the method of healthcare suppliers practice cultured equipment’s to increase awareness from their clinical and other data repositories and make a declared conclusion. Big data healthcare analytics has five processes: Data Acquisition, Data Storage, Data Management, Data Analytics, and Data Visualization & Report. The Figure 4 presents the process of big data analysis in healthcare management.
4.1. Data Acquisition

The big data in healthcare can be in a format of the structured, semi-structured or unstructured [40] and can be acquired from primary sources (e.g., CPOE, clinical decision support systems, electronic health records etc.) and secondary sources (laboratories, insurance companies, government sources, pharmacies, & HMOs, etc.) [6] the following are the most important sources of big data in healthcare.

Electronic Health Records: During the earlier time span the digitization of healthcare histories have delivered a basis for hospitals on medical datasets [41]. Electronic Healthcare Records data from Physician Notes, Lab report, ECG, Scan, X-RAY, Health sensor devices details, Medical prescriptions etc. These set of data are the foundation for personalized medicine and large cohort studies [42] for hospitals.

Image Processing: Medical pictures are one of the cause of data for analysis. Computed tomography (CT), photo acoustic imaging, ultrasound, molecular imaging magnetic resonance imaging (MRI), mammography fluoroscopy, positron emission tomography-computed tomography (PET-CT), and X-ray are some of the examples of imaging procedures that are well recognized within clinical settings [43].

Social Media: Healthcare data can be collected from social media like Facebook, twitter, LinkedIn etc. Social Medical logs data are usually used for analyzing disease spreading/transmission [44]. For example collecting information about a particular flu affected people from Twitter is faster than traditional method. Social networking cites like patients like me (www.patientslikeme.com) has more than 200,000 patients and is tracking more than 1,800 diseases [45].

Smart phones: Apps are the most important sources of data in the area of health care self-management. Nowadays, smart phones have health related apps like pedometers, fit bit that produces lot of data from number of stairs climbed, steps walked, and calories burned. Another app Mood panda is used to measure the individual mood and also anything from mental, emotional, and physical to social and environmental aspects of daily life. We also have apps for diabetes management I BG Star to monitor the blood glucose system. These apps create lot of data every day which contributes to healthcare research. Van Heerden et al. [46] used mobile phones to collect maternal health information from HIV-Positive Pregnant Woman in South Africa. Zhang et al. [47] showed that smart phones can be effectively utilized for domestic data gathering on infant feeding in rural china.

Web base data: The websites are also one of the most important source of healthcare data. The popular websites proving healthcare data are 23andMe and uBiome. 23andMe is a DNA study service providing evidence and implements for individuals to learn about and explore their DNA (https://www.23andme.com). uBiome is a Microbiome sequencing service that offers facts and tools for you to discover your micro biome (http://ubiome.com) [48].

4.2. Data Storage

The storage plays an vital part in big data. As a size of data in the healthcare industry is increasing we need an efficient and large storage platform. With such large cloud is the most promising technology. Clouds provides elasticity and proficiencies for get into data, creating awareness, accelerating the potential for scalable analytics solutions and driving value. Cloud computing is a powerful and promising technology to store enormous scale of data and perform enormous-scale and complex computing. It eradicates the need to sustain costly computing hardware, software and dedicated space.

Using cloud to examine big data in healthcare makes sense because:

Investments in big data investigation can be important and drive a need for efficient, cost-effective infrastructure.

Big data in healthcare organisations is a mix of internal and external foundations as they frequently keep patient’s most delicate data in-house, enormous volumes of big data generated by public providers and third-party may be located outside.

Hashem et al. specified that cloud computing structure can assist as an actual platform to address the data storage essential to accomplish big data analysis [49].

Table 2 describes different platforms for cloud storage. It also describes their respective vendors and tools with purpose, advantages, disadvantages and application in healthcare.
4.3. Data Management

Data Management in healthcare includes organizing, cleaning, retrieval, data mining, and data governance. It also includes the method of validating whether there is some scrap data or any missing values. Such data needs to be removed [50]. It helps in risk assessment of patients, personalized discharge plan. Major data management tools are Apache Ambari and HCatalog. These tools are explained in detail in table 3. Data retrieval is a process of extracting file or valuable information from large healthcare databases. Big data analysis in healthcare frequently contains information recovery and data mining [51]. Wang et al. [52] mention that “information retrieval is the process of searching within large document collections, and in healthcare it mainly covers medical text retrieval and medical image retrieval. Data governance refers to overall management of security, integrity, usability and availability of the data employed in an enterprise. Maintaining confidentiality of individual patient records is very important in Healthcare management. This section provides information on key legislation to healthcare data accessibility and Government regulations intended to address health data privacy. The important data governance Act are HITECH, HIPAA, HDI, GINA and FOIA.

Health Information Technology for Economic and Clinical Health (HITECH) Act declared as portion of the American Recovery and Reinvestment Act of 2009, was engaged into law on February 17, 2009, to encourage the acceptance and significant use of health information knowledge. The implementation of HITECH Omnibus rule takes data management to a new level by mandating technological aspects of maintaining the patient record confidentiality. Subtitle D of the HITECH Act addresses the security and privacy concerns related with the electronic broadcast of health information, partly, through numerous requirements that reinforce the civil and illegal application of the HIPAA rules. HITECH involves all healthcare providers, to implement electronic healthcare system [53].

Health Insurance Portability and Accountability Act of 1996 (HIPAA) was considered to deliver secrecy of ethics to safeguard patients’ health record and other medical information related to hospitals, doctors, health plans and other healthcare workers. HIPAA Title II also known as the administrative simplification provision which leads to the U. S. Department of Health and Human services to create national standards for handling electronic healthcare communications.

The HIPAA regulation is to declare that individuals’ health information is appropriately secured however permitting the movement of health information required to deliver and promote high value of health care and to safeguard the populace's health and subsequent regulations direct covered entities (e.g., a healthcare institutions) in the safety of

---

**Table 2. Cloud Storage Platforms in Healthcare.**

<table>
<thead>
<tr>
<th>Types &amp; Reference</th>
<th>Vendors</th>
<th>Components /platforms</th>
<th>Purpose</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big data Storage [49]</td>
<td>Google, Microsoft, Amazon, IBM</td>
<td>Cloud Services Azure S3, SmartCloud</td>
<td>The main purpose of is for saving large data and making easily accessible.</td>
<td>Enables the allocation of EHRs between certified physicians and hospitals in numerous Geographical areas, providing further appropriate contact to life-saving information and falling the need for replica testing.</td>
<td>1. Internet Bandwidth can be constraint if the speed is low. It may take longer time to upload files and retrieve them. 2. If no Internet, then no access to the storage 3. Data Security and Privacy is still a concern.</td>
<td>Cloud deployments for clinical applications with isolated or hybrid clouds provided these uses get maximum level of security, privacy, and availability. Nonclinical applications are a better fit for public arrangements but still must be wisely assessed.</td>
</tr>
<tr>
<td>Relational Databases [46]</td>
<td>Google, Microsoft, Amazon, Cloudera, IBM</td>
<td>Cloud SQL, SQL Azure MySQL or Oracle MySQL, Oracle, PostgreSQL, dashDB</td>
<td>They are fully succeeded SQL database service that makes it comfortable to set-up, maintain, bring about and administer relational MySQL databases in the cloud</td>
<td>Can access a familiar, highly available SQL database from Mapreduce applications, without having to worry about provisioning, management, and integration with other services.</td>
<td>There are connection, size, and App Engine specific limits in place.</td>
<td>Applications successively on Google App Engine or Google Compute Engine</td>
</tr>
<tr>
<td>NoSQL (Not only SQL) Database [7, 39, 49]</td>
<td>Google, Microsoft, Amazon, Cloudera, IBM</td>
<td>AppEngine Datastore, Table storage DynamoDB, Apache Accumulo, DB2</td>
<td>speedy and elastic NoSQL database facilitates for all application need reliable, single-digit millisecond potential at any scale.</td>
<td>1. Flexible data models offered by NoSQL databases allow unstructured or semi-structured data to be stored easily 2. NoSQL databases are based on horizontal scalability which allows easy and automatic scaling</td>
<td>1. NoSQL database experts are difficult to get since the application itself is relatively new. 2. Though it is claimed that less administration is required, but they still require significant level of skill and effort to install and maintain.</td>
<td>Since the significance of EHR application for stability of care and complete health systems, using NoSQL databases have important possible to lead to superior EHR applications in terms of scaling, flexibility and greater obtainability.</td>
</tr>
</tbody>
</table>
individually recognizable secure the health information in research [54].

Department of Health and Human Services (HHS) under the Health Data Initiative (HDI) proposed a reliable privacy-preserved medical recommendation system. In this medical system, the patients can contribute their secured ratings of the physicians on different health conditions based on their satisfactions [55].


4.4. Data Analytics

Data analytics is a process of transforming the raw data into information. Big data analytics in healthcare is classified into Descriptive, Diagnostic, Predictive, and Prescriptive Analytics [39]. The present research in academia and industry displays that retailers can attain up to 15 to 20% growth in ROI by setting big data into analytics [56].

Descriptive Analytics: It looks at past performance based on historical data. Descriptive Analytics is also known as unsupervised learning. It summarizes,

- What happened in the healthcare management?
- What is the impact of a parameter on the system?
- Diagnostic Analysis: Using historical data predicts the root cause of problem and diagnose
- Why did it happen?

Predictive Analytics: It analyzes both real-time and historical data, also known as supervised learning. It can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature. It cannot predict the future. It anticipate

- What will happen? What are the future trends?
- What is the decision based on past history?

Prescriptive Analytics: This analytics automatically synthesizes big data and provide advice on number of different possible outcomes before the decisions are actually made. The decision maker can take this information and execute. Prescriptive analytics is advanced than the descriptive and predictive analytics. It prescribes

- What should we do?
- What is the best outcome and how can we make it happen?

Table 3 describes the tools which are being used to manage healthcare data. There are many analytical tools to analyze the data. We have classified the tools into 8 different layers. Each layer is further subdivided into components.

<table>
<thead>
<tr>
<th>Layers</th>
<th>Components</th>
<th>Description</th>
<th>Web Link</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical Layer</td>
<td>Pentaho</td>
<td>Pentaho is use to design, develop and deploy Hadoop analytics solutions</td>
<td><a href="http://www.pentaho.com">http://www.pentaho.com</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apache Chukwa</td>
<td>Apache Chukwa is a data grouping method for handling huge circulated systems.</td>
<td><a href="http://chukwa.apache.org/">http://chukwa.apache.org/</a></td>
<td>[49]</td>
</tr>
<tr>
<td>Analysis Layer</td>
<td>Apache Pig</td>
<td>Apache Pig is a data-flow language in high-level, implementation of structure on behalf of similar calculation.</td>
<td><a href="https://pig.apache.org/">https://pig.apache.org/</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apache Spark</td>
<td>Apache Spark is use for processing large-scale data.</td>
<td><a href="http://spark.apache.org/">http://spark.apache.org/</a></td>
<td>[49]</td>
</tr>
<tr>
<td></td>
<td>Cloudera Impala</td>
<td>Cloudera Impala is similar handling (MPP) SQL query engine for data storage in a computer cluster</td>
<td><a href="http://www.cloudera.com/">http://www.cloudera.com/</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apache Tez</td>
<td>Apache Tez is a general data-flow programe structure, built on Hadoop</td>
<td><a href="http://tez.apache.org/">http://tez.apache.org/</a></td>
<td></td>
</tr>
<tr>
<td>Abstraction Layer</td>
<td>Apache Oozie</td>
<td>Oozie is a workflow processing structure that lets users describes a sequence of works inscribed in different languages then wisely linked them to each other.</td>
<td><a href="http://oozie.apache.org/">http://oozie.apache.org/</a></td>
<td>[7]</td>
</tr>
<tr>
<td>Processing Layer</td>
<td>Apache Cascading</td>
<td>Apache Cascading is use to generate and perform composite data processing workflows on a Hadoop cluster</td>
<td><a href="http://www.cascading.org/">http://www.cascading.org/</a></td>
<td>[24]</td>
</tr>
<tr>
<td></td>
<td>Hadoop Mapreduce</td>
<td>A YARN-based system is processed parallel on huge data sets</td>
<td><a href="https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html">https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html</a></td>
<td>[7-34-20-39-57]</td>
</tr>
<tr>
<td></td>
<td>Bidoop</td>
<td>Bidoop helps to analyze data sources and develop new analytical models</td>
<td><a href="https://www.mapr.com/apps/bidooop-layer">https://www.mapr.com/apps/bidooop-layer</a></td>
<td></td>
</tr>
<tr>
<td>Integration layer</td>
<td>Apache Flume</td>
<td>Apache Flume is use for mix together, combination, and moving huge volumes of log data</td>
<td><a href="https://flume.apache.org/">https://flume.apache.org/</a></td>
<td>[50]</td>
</tr>
<tr>
<td>(ETL)</td>
<td>Apache Avro</td>
<td>Apache Avro is serialization framework</td>
<td><a href="https://avro.apache.org/">https://avro.apache.org/</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apache Sqoop</td>
<td>Apache Sqoop is a transitional framework for data movement from non-Hadoop data into Hadoop data</td>
<td><a href="http://sqoop.apache.org/">http://sqoop.apache.org/</a></td>
<td></td>
</tr>
<tr>
<td>Managemen t Layer</td>
<td>Apache Ambari</td>
<td>Apache Ambari is a web-based tool for managing, monitoring, and provisioning Apache Hadoop clusters</td>
<td><a href="https://ambari.apache.org/">https://ambari.apache.org/</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apache H Catalog</td>
<td>H Catalog is a centralized sharing service and metadata management</td>
<td><a href="https://wiki.apache.org/confluence/display/Hive/HCatalog+UsingHCat">https://wiki.apache.org/confluence/display/Hive/HCatalog+UsingHCat</a></td>
<td></td>
</tr>
<tr>
<td>Coordinatio</td>
<td>Apache</td>
<td>A greater-presentation which coordinates service for distributed</td>
<td><a href="http://zookeeper.apache.org/">http://zookeeper.apache.org/</a></td>
<td></td>
</tr>
</tbody>
</table>
layers

Components

Description

Web Link

$n$ Layer

Zookeeper

applications

https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Storage

Layer

Hadoop HDFS

Hadoop HDFS is a circulated file structure that delivers high-

amount of contact on application data

https://hive.apache.org/

Apache Hive

Apache Hive is a data granary software abilities for managing

and querying bulk datasets exist in circulated storage

http://hbase.apache.org/

Apache HBASE

Apache HBASE is a dispersed database that enhances designed

data storing for large tables

MongoDB

MongoDB is document-oriented database outlined

http://cassandra.apache.org/

Apache

Cassandra

Apache Cassandra is a scalable multi-master database with no

opinions of disaster

4.5. Data Visualization

Data visualization is presenting the analytic results of

healthcare data into pictorial or graphical format for

understanding complex data and better decision making. It

can be used to understand pattern and correlation among the

data. The table 4 explains the tools that are currently being

used for visualizing the big data in healthcare. The table

describes description and features of the tool with its

application in healthcare.

Table 4. Visualization tools used in healthcare.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
<th>Features</th>
<th>Applications</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>R is a programming language for advanced statistics and data visualization.</td>
<td>- graphical display, control and manipulation of data</td>
<td>-R is been use by pharmaceutical company where it is used to plan clinical</td>
<td><a href="https://www.r-project.org/">https://www.r-project.org/</a></td>
</tr>
<tr>
<td></td>
<td>Cytoscape is a common bioinformatics bundle for data integration and biological network visualization.</td>
<td>- Combined tools for instantaneous analysis</td>
<td>to forecast the finale day of the conclusions built on prearranged temporary</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Apps used for modelling advanced analysis</td>
<td>studies of the data.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- for problem domains</td>
<td>Vulnerability Gene --slim (InitGO) based on GO hierarchy structure and</td>
<td>Qijun Liu et al.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>graph operations.</td>
<td>Yan Jiang, Chao Xu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- use to develop a semiautomatic technique to attain intact GO and self-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>governing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- use for bioinformatics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graphviz is a way of representing basic facts as figures of abstract, graphs and networks,</td>
<td>- flexible, -scalable -user-friendly software model</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IBM Watson Analytics is a powerful analytics tool for cloud data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table contains 5 tools that can be used in future for visualizing healthcare data with their description, features and related link.

Table 5. Visualization tools can be used in healthcare.

<table>
<thead>
<tr>
<th>Tools</th>
<th>Description</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodebox</td>
<td>It's a tool used for generative creative designs</td>
<td>- Data have been imported in various formats such as excel and in addition to that it is animation capable</td>
</tr>
<tr>
<td>Flot</td>
<td>Flot is pure Javascript conspiracy collection for jQuery wit, attractive and collaborating, humble structures as focused.</td>
<td>- Helps in plotting categorical and textual data and usage of combinations for displaying elements in the same data series.</td>
</tr>
<tr>
<td>FF Chartwell</td>
<td>FF Chartwell evolution is a simple series of numerical into editable data concepts.</td>
<td>- Produces interactive visualizations with toggling series</td>
</tr>
<tr>
<td>Raphael</td>
<td>Raphael uses SVG and VML creating vector. A JavaScript library on the web for graphics, so that each representations formed is also a DOM object.</td>
<td>- Multi-chart skills</td>
</tr>
<tr>
<td>Cross filter</td>
<td>A JavaScript library that is skilled at controlling data sets by in excess of a million records. Discovering huge</td>
<td>- Discover huge multivariate datasets</td>
</tr>
</tbody>
</table>


5. Applications of Big Data in Healthcare

Big data can be applied in almost all the areas of healthcare management. The potential application areas are fraud detection, epidemic spread prediction, Omics, clinical outcome, medical device design, insurance industry, personalized patient care and manufacturing, and

<table>
<thead>
<tr>
<th>Tools</th>
<th>Description</th>
<th>Features/ Related Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Charts</td>
<td>Google Charts offers fluctuating from basic distribute plots to categorised tree maps, a selection of data visualization formats.</td>
<td>-Cross-browser compatibility</td>
</tr>
<tr>
<td></td>
<td>Nephi allows users for difficult enquiry of links, and more for a better understanding of data relationships, and to both visualize and discover data social networks.</td>
<td>-Deep data analysis to study relationships</td>
</tr>
<tr>
<td>Tableau Public</td>
<td>Tableau is a flexible-to-practice device for crafting, communicating, insert and data visualizations them on your website.</td>
<td>-Desktop presentation but accomplished graphics are kept on a public server</td>
</tr>
<tr>
<td>Quadrigram</td>
<td>Quadrigram allows users to create entirely personalised visualizations using own data and several mechanisms from a built-in library of entirety from charts and graphs to quadrification and loaded flow.</td>
<td>-Drafting the ideas and generating speed prototypes</td>
</tr>
<tr>
<td>Perfuse</td>
<td>Perfuse is a data visualization tool which had been utilized by the IBM Visual Communication</td>
<td>-Quick data processing by cloud-based computing system</td>
</tr>
<tr>
<td>Many Eyes</td>
<td>This tool offers a stand for generating a change of visualizations to demonstrate data point relationships, create line and stack graphs, compare sets of values, examine text or view the many quantities of a total in a pie chart or tree map.</td>
<td>-Complete library of interactive visualizations</td>
</tr>
<tr>
<td>iCharts</td>
<td>iCharts is a web-based presentation skilled of generating fascinating data visualizations for the purpose of web. Integrated graphs and charts into a website or application or allocate concluded visualizations through the iCharts Chart Channel, or social media.</td>
<td>-Build animations, dashboards and more</td>
</tr>
<tr>
<td>Visualize free</td>
<td>Based on the marketable visualization device InetSoft, Visualize Free is a free alternate for scrutinizing over numerous data sets and alternatives to recognise developments and operate data with a few simple clicks.</td>
<td>-Create interactive, employable charts</td>
</tr>
</tbody>
</table>

Table 6 describes all the tools that need programming knowledge to visualize healthcare data.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Features/ Related Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3. js</td>
<td>D3 is javascript library for creation of data visualizations.</td>
<td>- Creates interactive SVG bar charts and bind arbitrary data to DOM</td>
</tr>
<tr>
<td>Flare</td>
<td>An Action Script collection aimed at producing data visualizations that runs in Adobe Flash Player.</td>
<td>- Skilled of cooperating, difficult graphics</td>
</tr>
<tr>
<td>SAS Visual Analytics</td>
<td>SAS Visual Analytics is the instrument to discover data sets of complete dimensions on visual basis for further widespread analytics.</td>
<td>- Maintenances visual encoding, data management, interaction and animation</td>
</tr>
<tr>
<td>Network X</td>
<td>Network X is skilled enough for producing graphs, digraphs and multigraphs based on datasets involved for multiple media formats. It based on the Python programming language</td>
<td>-Study the organization, dynamics and purposes of difficult networks</td>
</tr>
</tbody>
</table>

pharmaceutical development etc. [20]. Moreover the application of big data is widely adopted in personalized healthcare which offers an individual centric approach [58].

5.1. Applications of Big Data in ‘Omics’

“Omics” data refer to significant datasets in the organic and molecular fields (e.g., proteomics, metabolomics, macrobiotics, genomics etc.). Application of big data on this study is to realize the strategies of diseases and increase the specification of medical treatments (e.g. “precision medicine”) [59]. With the advance in metabolomics, proteomics, genomics, and other types of omics know-hows through the previous eras, a remarkable volume of data associated to molecular biology has been formed [60].

Genomics is the study of genes and their functions [61]. Application of big data in genomics will help to prevent or cure diseases and delivering personalized care to each patient [62]. This area is in still emerging period with presentations in particular concentrated regions, for example leukemia, diabetes, and cancer [63-64]. Pathway analysis is mostly used for high-quality of genome-scale data [65], there are three generations of same structures used in pathway analysis [66]. The first generation tools are Clue Go, Onto-Express and GoMiner [67]. The most popular tool for second generation is GSEA [68], and the example for the third generation tool is Pathway-Express [69].

Proteomics is the study of proteome on their structures and functions. A proteome is the entire set of proteins in a cell. ExPASy (http://www.expasy.org/proteomics) lists dozens of databases on proteomics and over 100 tools. Big data application in proteomics, will have a major role in predicting and preventing human cancer [70]. Find Mod [71] and CSS-Palm [72] are frequently used for PTMs prediction.

Metabolomics is the systematic concept of chemical procedures including metabolites. The database BiGG used Genomic-based reconstruction of human metabolism for systems biology [73].

5.2. Insurance Industry / Payer

Healthcare Insurance companies/ payers are using big data in underwriting, fraud deduction, and claim management. Insurance providers are observing further than algorithmic fraud revealing practices that are claim-centric, to ones that are person-centric [74]. For example how many related claims were been submitted by the same personality or stated the identical treatment in different insurance companies.

5.3. Medical Device Design and Manufacturing

Big Data implement facilitates a wider set of device materials, delivery methods, and tissue interactions, anatomical configurations to be evaluated. Calculation techniques and Big Data can plays a significant role in medical system strategy and manufacturing [75].

5.4. Pharmaceuticals

Big data is used during all phases of pharmaceutical development, particularly for drug discovery [57]. Pfizer has recently initiated Precision Medicine Analytics Environment program that associates the dots among electronic medical record data, clinical trial, and genomic to identify chances to rapidly convey innovative medicines for particular patient populations.

5.5. Personalized Patient Care Healthcare

Big Data will make possible to bring best and modified patient care. In nearby future, fresh big data-derived influences will prompt suitable updates of diagnostic assistance, clinical guidelines and patient triage to permit more particular and modified treatment to advance medical result for patients (Yang et al., 2014)

6. Conclusions and Future Research Directions

This study reviewed the literatures on big data in healthcare. The works gap in the terms of method is recognized and connected with the empirical study.

The most challenge parts for big data in healthcare are data privacy, data leakage, data security, efficient handling of large volumes of medical imaging data, information confidentiality and security, wrong use of health data or failure to safegourd the healthcare information, and understanding unstructured clinical notes in the right context, extracting potentially useful information [76].

Some limitations have been identified after review, firstly there are limited publication that are available on big data in healthcare management.

This reading addressed the major experiments on data governance in healthcare. More detail study on data governance component can be future research topic.

The authors suggest few new data visualization tools to the healthcare analyst to make effective decision making. Big data has a great prospective to progress healthcare management and transform healthcare industry to next level. The review article will be benefiting the healthcare academicians, practitioners, researchers who are engaged in the areas of healthcare Management.

Conflict of Interests

The authors declared that there is no conflict of interests about the publication of this paper.

Acknowledgements

The authors are grateful to the editor and anonymous reviewers for those constructive comments and recommendations.
References


