Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 1, April 2017.

BIG DATA IN HEALTHCARE: REVIEW AND OPEN RESEARCH ISSUES

Mohammad Ashraf Ottom

(Received: 18-Oct.-2016, Revised: 29-Dec.-2016, Accepted: 21-Jan.-2017)

ABSTRACT

The globe is generating a high volume of data in all domains, such as social media, industries, stock markets and healthcare systems. Most of data volume has been generated in the past two years. This massive amount of data can bring benefits and draw knowledge to individuals, governments and industries and assist in decision making. In healthcare, an enormous volume of data is generated from healthcare providers and stored in digital systems. Hence, data are more accessible for reference and future use. The ultimate vision for working with health big data is to support the process of improving the quality of service in healthcare providers, reducing medical mistakes and providing a promoting consultation in addition to providing answers when needed. This paper provides a critical review of some applications of big data in healthcare, such as the flu-prediction project by the Institute of Cognitive Sciences, which combines social media data with governmental data. The project aim is to provide swift response about flu-related questions. The project should study human multi-modal representations, such as text, voice and images. Moreover, integrating social media data with governmental health data could create some challenges, because governmental health data are considered as more accurate than subjective opinions on social media. Another attempt to utilize big data in healthcare is Google Flu Trends GFT. GFT collects search queries from users to predict flu activity and outbreak. GFT performed well for the first two to three years; however, it started to perform worse since 2011 due to people behaviour changes. GFT did not update the prediction model based on new data released by the Centre for Disease Control and Prevention-US (CDC). On the other hand, ARGO (Auto Regression with Google) performed better than all previously available influenza models, because it adjusts people behaviour changes and relies on current publicly available data from google-search and CDC. This research also describes, analayzes and reflects the value of big data in healthcare. Big data has been introduced and defined based on the most agreed terms. The paper also explains big data revenue forecast for the year 2017 and historical revenue in three main domains: services, hardware and software. Big data management cycle has been reviewed and the main aspects of big data in healthcare (volume, velocity, variety and veracity) have been discussed. Finally, a discussion has been made of some challenges that face individuals and organizations in the process of utilizing big data in healthcare, such as data ownership, privacy, security, clinical data linkage, storage and processing issues and skills requirements.

KEYWORDS

Big data, Healthcare, Hadoop, Google flu trends, Big data challenges, Cycle of big data management.

1. INTRODUCTION

Some people could be amazed about the fact that 90% of the current volume of data have been generated in the past two years and that the amount of data is expected to grow 40% per annum according to Aureus Analytics [1]. Big data is a vast and vague terminology that carries many meanings depending on time and technology generation. For example, one megabyte of data in the 1950s was considered as a big amount of data when IBM manufactured the disk file IBM-350 with a total capacity of a couple of megabytes [2], where the US presidential debate between Barack Obama and Mitt Romney in 2012 has produced about 10.3 million tweets in ninety minutes [3]. In addition, other social media and business transactions produce an

A. Ottom is with the Computer Information Systems Department, Yarmouk University, Irbid, Jordan.
E-mail: ottom.ma@yu.edu.jo.

enormous amount of data in every minute. Therefore, the term big data is a changing term depending on time, since what we consider as "big" in the present may not be big in the future.

The main objective of this manuscript is to describe a current issue in Information Technology research, which is big data in healthcare field. This manuscript is to provide insight into values and opportunities of big data in healthcare industry. It also attempts to draw some attention to market revenues and economic impacts of big data. Some big data applications in healthcare are reported and commented. Some related research issues are identified and discussed. In addition, the paper criticizes some applications of big data in healthcare, such as the flu-prediction project by the Institute of Cognitive Sciences, Google Flu Trends and Auto Regression with Google (ARGO).

A study [4] investigated the term big data to establish a solid definition that describes what is actually meant by big data. The study found that big data is more frequent in four domains: information, technologies, methods and impacts. The study proposed the most prominent definition: “Big data is the information assets characterized by such high volume, velocity and variety to require specific technology and analytical methods for their transformation into value”. In information technology, big data could be a huge amount of data that is beyond the traditional database capabilities [5]-[6].

Several reports claim that every day the globe produces 2.2 terabytes of data in which only 10% are structured data, where 90% are unstructured data. Further, 90% of data have been generated in the past two years [1]. Nowadays, organizations are holding and collecting a huge amount of data, but do not know what they have and how to utilize the collected data properly. This is like a person who has a big library at home or in his office, but never read the books. However, recently the term big data is used to represent the huge data that people generate around the globe with a little debate about its importance. Big data has created some controversy among researchers about the data significance and future importance. Some supporters claim that big data could be the new trend for innovation and discovering knowledge from outsized amounts of data in the near future and will lead to the evolvement of a new the science in parallel with computer machine learning and data mining advancement [7]-[8]. On the other hand, some disputed big data importance and stated that bigger data is not always better [9]. Michael Jordan in an interview with IEEE spectrum stated about big data: “It is like having billions of monkeys typing. One of them will write Shakespeare”. However, he admitted that data analysis can produce inferences about a certain problem, but with a certain level of quality and with an error margin [10]. Big data also raises an argument about the future of data mining in big data era and the possibility of big data to replace data mining completely [7]. This argument needs in-depth future research to let see what the coming days carry for data mining.

Despite the debate created around the term big data, a recent study [11] showed the importance and market revenue of big data vendors, such as HP, IBM, DELL, SAP, ORACLE and SAS in three major domains: hardware, software and services. The study found that the total revenue reached US$ 18.6 billion in 2013, where 40% of the revenues were in the service domain, 38% in the hardware domain and 22% in the software domain. Figure 1 shows big data revenues by domain for the year 2013.

The study [11] also forecasted the market revenues for the year 2017 and compared the revenues for the years from 2011 until 2017. The market revenue for 2011 was US$ 7.3 billion and is expected to reach $US50.1 billion in 2017. Figure 2 shows the market revenues for the years from 2011 to 2017.

Another study [12] by the International Data Corporation (IDC) about big data storage forecast for the period from 2014 to 2019 found that hardware platforms and systems are estimated to grow at a Compound Annual Growth Rate (CAGR) of 20.4% from 2014 to 2019.

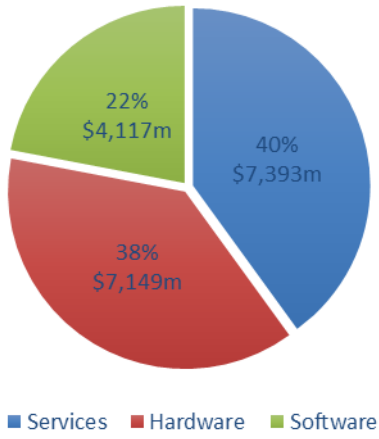

Figure 1. Big data revenue by domain in 2013 in US$ million.

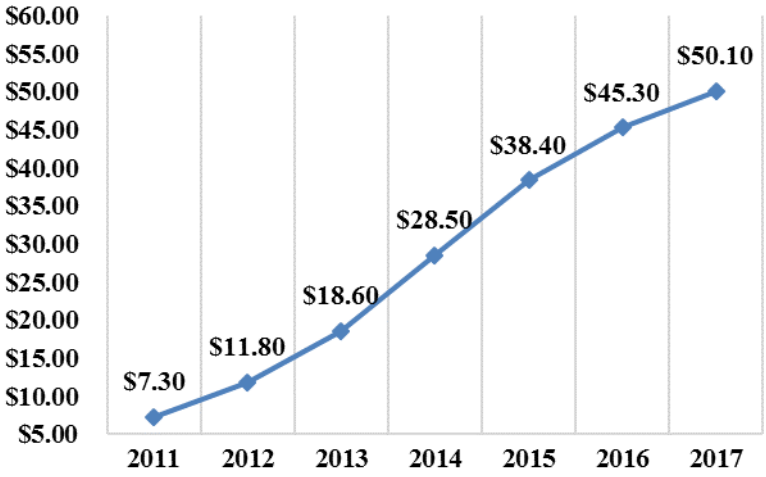

Figure 2. Big data market forecast between 2011 and 2017 in US$ billion.

In healthcare domain, the advancement in information technology and the capability of storing more data in the digital age have driven countries and governmental organizations to computerize health records and produced what we call Electronic Health Record (EHR) or Electronic Medical Record (EMR). EHR/EMR is the electronic form of a patient's medical history that is equivalent to the traditional hard copy of the patient's medical record. EHR system enables patient's records to be shared across healthcare providers in a certain state or globally. EHR records may include a range of data, including general medical records, family medical history, patient examinations, patient treatments, allergies, immunization status, laboratory results and radiology images [13]. The adoption of EHR in many countries has encouraged healthcare providers to store patients' information in an electronic form which produced a large amount of structured and unstructured data. For instance, reports about the United States of America stated that U.S EHR reached 150 exabytes by the end of 2011 and may soon reach the zettabyte and yottabyte era [14]. Another report [15] analyzed the future of EHR market. In terms of geographical analysis, the report determined the market trends for five regions (Europe, Asia and the Pacific, North America, Latin America and the rest of the world). The report concluded that EHR market was valued at US$ 18.9321 billion in 2014 and predicted a compound annual growth rate of 5.4% for the period from 2014 to 2023 that may reach the peak of US$ 30.280 billion by 2023.

2. BIG DATA ASPECTS IN HEALTHCARE

The literature shows that the initial aspect for big data was introduced by Doug Laney in 2001 [1]. Laney argued that e-business and e-commerce system transactions are generating a high volume of information, rapid growth of data (velocity) and different types of data from several domains (variety). In addition, enterprises started to consider data as an asset they became keener to store data for future use. Volume, velocity and variety formed the basic aspects of big data called the 3Vs. Figure 3 shows the 3Vs big data model.

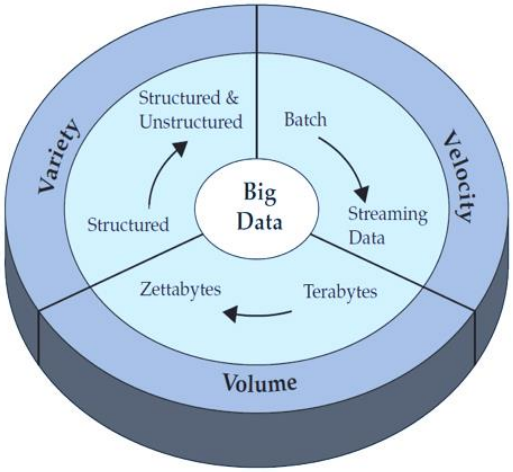

Figure 3. IBM big data model characterized by volume, velocity and variety [16].

Volume refers to dimension, quantity or magnitude of data. This aspect of big data is changing with time. For example, one terabyte of data could be a large amount of data today. However, due to rapid development of computer storage, this couldn't be considered as large data in the future. IBM conducted a survey to determine how big the database should be to be considered as big data. It was found that one terabyte dataset can be considered as big [17], but one terabyte of data may not be considered as big data in the future as the world is shifting to zettabyte era. Variety refers to data diversity from different sources, where data can be structured or unstructured. Velocity refers to data transfer in terms of speed and contents change [18].

In healthcare, big data aspects can be categorized into 4Vs: volume, variety, velocity and veracity [19], as shown in Figure 4. Volume, variety and velocity are identical to the general aspect of big data that shown in Figure 3. Though, an additional aspect has been added, which is veracity. Due to large volume of data and variety of sources, healthcare data varies in quality and complexity. Usually, healthcare data contains biases, missing feature values and noise, which could affect the decision-making process. Additionally, reliable data could reduce the cost of data processing [19]. Mainly, there are two types of data quality problems. The first type is attributed to technical matters, whereas the second type is the truthfulness of data and data sources [20].

Big data problem solving procedure consists of several phases. First comes the process of capturing and collecting data from one source or from different sources. Healthcare data volume is huge and derived from different sources, such as hospitals, medical centres, pharmacies, pathologies …etc. The process of capturing healthcare data can generate some challenges, such as data characteristics, heterogeneity, storage capacity, storage medium, cloud storage, compression tools and plans for backup and disaster recovery. Thus, IT industries, such as Oracle, have developed solutions to handle a very high transaction volume in a distributed environment and provide support to the research community in the field [21]. Similarly, HDFS (Hadoop File System) can handle a huge volume of data across multiple machines or distributed

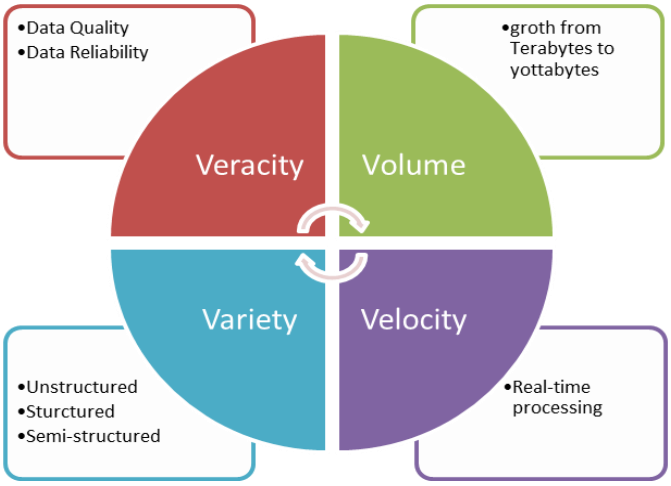

Figure 4. Big data in healthcare (4Vs).

systems [22]. Second comes the process of organizing data. In some cases, data need to be transformed from its original state into a new state to prepare data for analysis using big data solutions. In healthcare, migrating data from legacy database management systems to Hadoop system, for example, require to reformat data into a more beneficial structure, such as hierarchical structure [23]. The third phase is data integration from different sources. In healthcare, Hadoop and NoSQL solutions support data integration from data warehouses, hospitals and social platforms like Facebook …etc. [24]. Fourthly, the purposes of analysis should be identified. For example these purposes could be, obtaining a valuable insight from huge data, predicting healthcare fraud, identifying a useful pattern, predicting patients' behaviour, detecting diseases at earlier stages or tracking disease outbreaks and transmission …etc. [25]. Big data can be analyzed with advanced software tools used in the analytic field, such as online analytical processing, data mining, statistical analysis, machine learning …etc. Finally, decision makers can act based on the findings of the analysis process. Figure 5 shows the cycle for big data management [26].

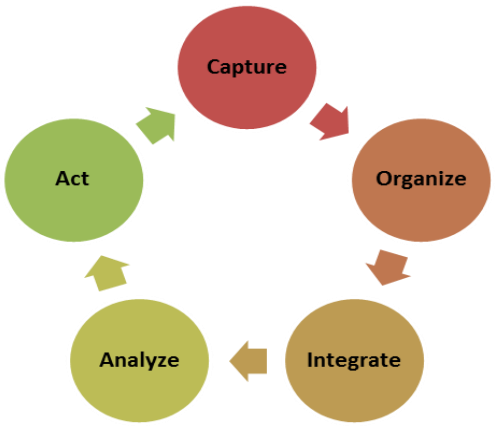

Figure 5. The cycle for big data management.

3. BIG DATA IN HEALTHCARE

The main objective of big data is to gain knowledge from a high volume of unstructured data that comes from numerous sources. Health data and computational techniques can be used to answer clinical questions. For instance, flu-prediction project by the Institute of Cognitive Sciences at Osnabrück University is combining social media data, e.g. Twitter, with CDC data

(Centre for Disease Control and Prevention in the USA). Flu-prediction project [27] gives the users the opportunity to ask questions about flu-related subjects and provides answers in seconds. However, the precision of flu-prediction project answers needs more enhancement and development. Besides, combining social media data with governmental health data will possibly create some challenges, because governmental health data is considered as more accurate than subjective opinions on social media.

Another attempt on big data in healthcare is Google-Flu Trends. Google Flu Trends (GFT) is a web service operated by Google. The main aim is to provide some information on influenza activity for selected countries (more than 25 countries). Google Flu Trends collects search queries from users, then it tries to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu [28]. GFT performed well for the first two to three years. Unfortunately, Google Flu Trends started to perform worse since 2011 and missed the peak in year 2013 by 140 percent according to a research team from Northeastern University, the University of Houston and Harvard University [29]. However, GFT failure doesn't reduce the importance and value of big data. Further, it indicated the possibility of using big data in the healthcare sector and opened the door for more future research such as Auto Regression with GOogle search data (ARGO).

A research team from the Department of Statistics at Harvard University investigated the failure of Google Flu Trends and proposed a new model for influenza tracking called ARGO (Auto Regression with GOogle search data) [30]. They found the main causes of Google Flu Trends failure. First comes, the lack of ability to adapt with people behaviour changes online. For example, people usually search for influenza symptoms and diagnostics, but lately, people behaviour shifted to search for news about the influenza season. In addition, different keywords are used for searching. Secondly, Google Flu Trends didn't utilize the new data released by CDC (Centre for Disease Control and Prevention) which may allow them to adjust and enhance the model using more reliable data [31].

ARGO performed better than all previously available Google-search–based tracking models for influenza, as well as the Google Flu Trends. ARGO utilizes the current publicly available data from google-search, data collected from Google Flu Trends and other publicly available data from other vendors, such as CDC. The main features of ARGO as described by the research team [30] were the ability to integrate influenza data with other data related to epidemics and monitor and record the changes of people search online behaviour. Most important, ARGO can be used for real-time tracking of other social events and diseases. Despite the decent features of ARGO and its performance against previous models, AGRO is not guaranteed to work for ever or to success next year for example, because of public data and people search behaviour shift. Furthermore, any changes to the inner-works of the search engine or any changes in the way information is displayed to users will affect the accuracy of ARGO.

From Google Flu Trends GFT and ARGO, we may comprehend the significance of the work accomplished by GFT and the importance of the idea. Despite the failure of GFT, it draws a road map for others to investigate problems and find solutions and that what AGRO team has been doing. Further, we may expect other research teams to focus on all versions of flu-season-trends to find a better version that may in future become a product and to emphasize the importance of big data in healthcare and other sectors. A study on big data analysis [32] confirmed the importance of big data in modelling disease spread and real-time identification of emergencies. Another study [33] demonstrated that big data analysis can be used to discover disease patterns and record disease outbreaks over the world, especially when a swift information is needed. The study added that public health issues can be improved with the analytical approach, where a large amount of data can help determine the needs, offer required services and predict and prevent the future crises to benefit the people. Further, detecting fraud in healthcare becomes more efficiently. Not to forget is the most important role of big data analysis in healthcare, which is to enhance the quality of care and services for patients. In

addition, big data can help with knowledge distribution across healthcare providers in one country or across the world. Some countries have a lack in medical expertise. So, knowledge and guidance driven by big data may provide useful and rapid information for practitioners in regional and third world countries. For example, Swine Flu (H1N1) was first reported in April 2009 infecting millions of people and estimated deaths worldwide, due to H1N1, were about 18500 people according to WHO report [34]. When a certain country or organization has huge volume of health data about the generic influenza viruses, then big data tools could provide swift knowledge, such as initial diagnosis and treatment procedures. These outcomes and knowledge can be distributed and shared across healthcare providers worldwide to save lives and reduce the impacts before it is too late.

4. BIG DATA TOOLS IN HEALTHCARE

Mike Cafarella and Doug Cutting can be considered as the founders of Hadoop, which is an evolution from the open source web search engine (Apache Nutch). Apache Nutch creators understood the limitation of nutch and the difficulty to reach the very huge number of webpages in the internet. At the meantime, Google published a research paper about its own storage system architecture called Google's Distributed File System (GFS) which solved nutch storage needs and drew a map to implement Nutch Distributed File System (NDFS). It also opened the space for other research issues and development opportunities. Google also introduced MapReduce in another research paper and nutch team implemented MapReduce to nutch which later ran on MapReduce and NDFS. In 2006, nutch developers formed an independent project based on MapReduce and NDFS, called Hadoop, which is the elephant toy name for Doug Cutting's son [35].

Hadoop is an open source infrastructure software or a framework to store and process a huge dataset. It is an open source project under apache. Basically, Hadoop performs two fundamental operations; storing data and processing data. Hadoop stores information by using Hadoop Distributed File System (HDFS). HDFS uses the cluster architecture to store data across different machines (nodes), such as PCs and servers, which gives the ability to store a huge volume of data on thousands of nodes similar to what Yahoo is doing nowadays. The second operation that Hadoop preforms is data processing. Data processing is accomplished by another component called MapReduce. Processing data includes counting keywords in the dataset, aggregating data and searching in the data. The traditional architecture of processing data stored on different clusters was to move the data from the cluster to the software. This operation may cause bandwidth problems and consume a long time, especially with big data. However, MapReduce performs opposite to traditional architecture by moving the software process to the node (map) instead of moving the data to the software, and then collecting the answers or the process output (Reduce). This is where the name MapReduce comes from. Hadoop ecosystem contains different sub-projects to help give Hadoop more flexibility and ease of use, such as Pig and Hive [22], [36].

Pig is a high-level programming language for expressing a data analysis program that is designed to ease and utilize Hadoop programming and tasks for users with little programming skills, without having to know MapReduce coding to achieve a certain task. Further, Pig is an extendable language, where users can define their own functions. In addition, tasks in Pig are encoded to permit the system to optimize task execution automatically, which gives the users more room to focus on semantics rather than on efficiency [37]. Hive is a data warehouse infrastructure used for several operations on distributed storage and managing of large data using HiveQL (pretty similar to Structure Query Language SQL). Hive was developed by Facebook, then Apache Software Foundation used the Facebook version and enhanced it to form an open source tool under the name Apache Hive. The main idea behind Hive is to solve the complexity of MapReduce programming; therefore, instead of writing MapReduce program in Java, users are able to code a query for MapReduce job and process it to produce the same

result as in MapReduce [38]. HBase is a part of Hadoop ecosystem that runs on the top of Hadoop Distributed File System (HDFS). The main reason for developing HBase is that HDFS read/write of data is a sequential matter, where HBase allows random real-time access and stores data in a similar way as in the conventional RDBMS (columns and rows), where HDFS stores data as a collection of files. In addition, HBase provides faster access and operation on a huge volume of data. HBase has its own Java client API and tables in HBase can be used both as input/output for MapReduce tasks [36]. Figure 6 shows a common Hadoop ecosystem.

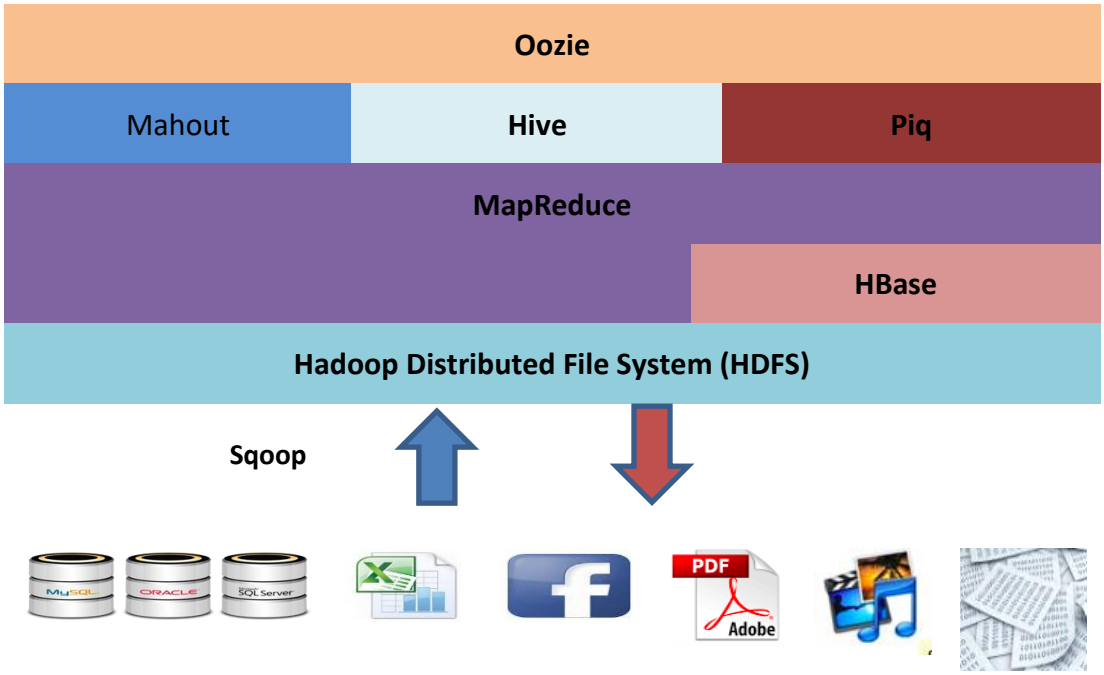

Figure 6. Hadoop ecosystem.

In big data analytics, data loading into clusters and processing occur in parallel with other data that resides on one location or is distributed over different locations. Loading a huge volume of data into Hadoop can be a difficult task, where accessing data directly by mapReduce may cause some complications, where Sqoop can facilitate the process of importing and exporting data, generate Java classes to allow interaction with data and allow users to import data directly into Hive data warehouse [39]-[40]. Mahout is an implementation of scalable machine learning algorithms by Apache Software Foundation. Mahout would be a good choice when data is distributed on several machines and when data volume is huge. In machine learning applications, larger data size could produce better accuracy. As the number of training data increases, Mahout's performance also increases [41].

Hadoop is considered as the best technology for big data in healthcare industry because of its ability to handle a huge volume of data in different formats, leaving no data behind, as stated by Charles Boicey; an information solutions architect at the University of California Irvine. Hadoop has been used widely in healthcare industry for cancer treatment and genomics, monitoring patients, clinical support network, healthcare intelligence, fraud prevention and detection …etc. [42].

The main reason why cancer has not been cured so far is that cancer transforms into diverse patterns based on every individual's genomics, and there are over three billion base pairs that constitute human DNA. In the future, big data tools such as Hadoop could provide the opportunity to store and map the three billion base pairs for every cancer patient. In monitoring

patients, Hadoop technology has been successfully used to monitor about 6200 children patients in Children's Healthcare of Atlanta (CHOA) in the United States. CHOA used sensors beside each child to track different signs, such as blood pressure, heartbeat and respiratory rate. Sensors collect a huge volume of data which is beyond legacy systems, then use the collected data successfully to produce patterns that alert medical staff to provide medical attention to patients when it is needed. Explorys is a company holding the largest database in healthcare. Explorys used Hadoop technology to analyze the huge volume of data and produced an analytical tool which assists medical staff to provide the most suitable treatment for individual patients or patient populations [42].

5. BIG DATA RESEARCH CHALLENGES IN HEALTHCARE

Big data can bring huge benefits to individuals, organizations, countries and to the world. However, benefits can bring risks as well, such as the lack of privacy and security. Further, many big data tools are open source and free to use tools, which may cause back-doors for intrusions, hackers and data theft. Hence, confidentiality, security, integrity and availability should be measured.

Privacy and Security: The first concern when working with big data is privacy and security. Privacy and security are a key concern for individuals and organizations that hold information/data about people, products, transactions …etc. Health data obtained by healthcare providers and medical practitioners from individuals' may contain private and confidential data. Individuals data must be handled with enormous care to protect people's privacy and confidentiality. Some approaches used to enhance the security level and obtain some confidentiality are [43]: First, individual identification is deleted during data collection (anonymous data). Second, individual identification is recorded initially during data collection and then removed. In this type of identification, there is a chance to re-identify the patient because patient information has been recorded at some stage (anonymized data). However, the removal of personal health data requires the removal of data elements in Table 1. Removing these data to meet the de-identification act [44] can affect the outcome of data analysis [45]. Third, encoding and encrypting data, however, give some chances to identify the encryption key using the advancement of computer technology which still exists. Privacy advocates and data regulators are gradually complaining about data collection and data usage in big data era, calling for a sophisticated protocol that achieves balance between individual privacy and research benefits [46].

Data Ownership: Data ownership describes a critical and ongoing challenge in big data applications in healthcare and other fields. Though petabytes of health data reside in healthcare providers' premises and governmental healthcare systems, these are not really owned by them. On the other hand, patients believe that they own the data. This controversy maybe ended in the legal system to resolve the ownership issues unless healthcare providers receive a written consent from patients prior to utilizing data for commercial or research purposes [47]-[48].

Clinical Data Linkage (CDL): CDL defines one of big data issues, where there are two or more tuples or instances relating to the same individual or entity. CDL occurs in the healthcare data when one individual or entity is stored on multiple healthcare providers such as two different hospitals. In addition, CDL can be quite challenging if the infrastructure of healthcare system is heterogeneous. In Australia, for example, clinical data linkage amongst patients with Spinal Cord Injury (SCI) is a challenge, because the Australian healthcare system is distributed over different states, and there is lack of coordination between healthcare providers in the states and in the federation. Hence, duplication exists and patient information that exists on many different sources is inconsistently updated [49].

Data Characteristics: Big data are usually collected from different sources and in different formats, which make them heterogeneous data. This may cause a limited value for big data,

since data may be incomplete, poorly described, improperly collected or outdated [50]. Further, the main purpose of data collection in healthcare systems is not for data analytics. Usually, healthcare data analytics is a secondary purpose of data collection. For instance, patients' data is collected primarily for payment, as well as tracking patient progress, treatment and clinical status. When the data is used for knowledge discovery, it may compromise the reliability and validity of any resulting models, because data has been collected for a different purpose. Hence, this creates a challenge for big data analysis and needs more effort to ensure data reliability [45].

Storage and Processing Issues: Doubtlessly, there is a significant advancement in computer storage technology. However, data grow significantly whenever a new storage technology is invented due to the huge amount of data collected/transferred by social media, healthcare providers, business transactions, stock markets …etc. The enormous amount of data created and generated around the globe put great pressure on data processing. The processing issue could be solved by bringing the software to data instead of sending data to the software or by transmitting only data which is important to the analysis process. However, this may create problems on data integrity and data source [47]. Health data movement is another related challenge; for instance, how huge volumes of data sized as giant (petabytes) move from data centres to the cloud. In this case, huge volumes of data are transported physically using physical media, such as FedEx and Amazon Snowball [51], but this could create another challenge when data needs an update. Storage capacity is another challenge in healthcare big data. For instance, one single human genome sequencing requires about two terabytes of storage capacity [52].

Skills Requirement: Big data researchers have been engaged with plenty of studies and researches that describe big data applications and technology development in data storage and analysis. However, there has been little attention to skills required for individuals to works in the big data field. A recent study [53] investigated the required skills to deal with big data and concluded that a big data specialist should be a combination of computer scientist and statistician with significant industry knowledge. Further, experts cite major shortages in big data specialists [51]. Further, educational institutions are responsible for making students aware of the new trend in data analytics and the required skills and technology for the industry.

6. SUMMARY AND CONCLUSION

This paper criticizes some applications of big data in healthcare, such as flu-prediction project by the Institute of Cognitive Sciences, which combines social media data with governmental data to provide swift response about flu-related subjects. Regardless of great efforts made by the Institute of Cognitive Sciences, the project should study human multi-modal representations (verbal communication, emotions, text and images). Moreover, integrating social media data with governmental health data will possibly create some challenges, because governmental health data is considered as more accurate than subjective opinions on social media. Particularly, integration between different authoritative sources of data could enable the composition of two complementary points of view of the same problem, which could affect the outcome of knowledge acquisition. Google Flu Trends GFT collects search queries from users to predict flu activity and outbreak. GFT performed well for the first two to three years; however, it started to perform worse since 2011 due to people behaviour changes. GFT did not update the prediction model based on new data released by the Center for Disease Control and Prevention-US (CDC). On the other hand, ARGO performed better than all previously available influenza models, because it adjusts people behaviour changes and relies on current publicly available data from google-search and CDC.

This research paper also describes, analyzes and reflects the value of big data in healthcare. Big data has been introduced and defined based on the most agreed terms. The paper explains big data revenue forecast for the year 2017 and historical revenue in three main domains: services, hardware and software. Big data management cycle has been reviewed and the main aspects of

big data in healthcare (volume, velocity, variety and veracity) have been discussed. Finally, a discussion has been made of some challenges that face individuals and organizations in the process of utilizing big data in healthcare, such as data ownership, privacy and security, clinical data linkage, storage and processing issues and skills requirement. The paper found that each issue mentioned in the current work requires further research because of their importance, where this will be the future research, especially on privacy and security issue. The paper also concludes that the world is moving toward data-driven approach which relies on data to perform business decisions. It is also important to understand that failure can be the first step toward success. Hence, the failure of Google Flue Trends might be the first step toward a successful model to utilize big data successfully and efficiently in healthcare.

Table 1. De-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule, Washington, DC: Department of Health and Human Services [cited January 16, 2017].

(A) Names
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
(D) Telephone numbers
(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers
(M) Device identifiers and serial numbers
(F) Email addresses
(N) Web Universal Resource Locators (URLs)
(G) Social security numbers
(O) Internet Protocol (IP) addresses
(H) Medical record numbers
(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers
(Q) Full-face photographs and any comparable images
(J) Account numbers
(R) Any other unique identifying number, characteristic or code, except as permitted by paragraph (c) of this section; and
(K) Certificate/license numbers
(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is subject of the information.

REFERENCES

[1] R. Jayanthi, "Big Data Applications in Healthcare," in: Impact of Emerging Digital Technologies on Leadership in Global Business, USA: IGI Global, p. 202, 2016.

[2] K. W. Oh, P. Lee and Y. W. Choi, "Enhanced Unlatch Operation of Disk Drive for Low Temperature Environment," Procedia Eng., vol. 131, pp. 906–913, 2015.

[3] A. Li, "Presidential Debate Most-Tweeted Event in U.S. Political History," 2012, [Online], Available at: http://mashable.com/2012/10/04/presidential-debate-twitter/#p54DATw_Wuqz.

[4] A. De Mauro, M. Greco and M. Grimaldi, "What is Big Data? A Consensual Definition and a Review of Key Research Topics," AIP Conference Proceedings, vol. 1644, pp. 97–104, 2015.

[5] E. Dumbill, "Making sense of big data," Big Data, vol. 1, no. 1, pp. 1–2, 2013.

[6] D. Fisher, R. De Line, M. Czerwinski and S. Drucker, "Interactions With Big Data Analytics," Interactions, vol. 19, no. 3, pp. 50–59, 2012.

[7] C. Wu, R. Buyya and K. Ramamohanarao, "Big Data Analytics = Machine Learning + Cloud Computing," p. 27, Jan. 2016.

[8] N. Council, Frontiers in Massive Data Analysis, The National Academies Press Washington, DC, 2013.

[9] D. Boyd and K. Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological and Scholarly Phenomenon," Information, Commun. Soc., vol. 15, no. 5, pp. 662–679, 2012.

[10] L. Gomes, "Machine-learning Maestro Michael Jordan on the Delusions of Big Data and other Huge Engineering Efforts," IEEE Spectrum, vol. 20, Oct. 2014.

[11] J. Kelly, "Big Data Vendor Revenue and Market Forecast," Wikibon Artic. Febrero, 2014.

[12] Ashish Nadkarni, Iris Feng and Laura DuBois, "Worldwide Storage in Big Data Forecast, 2015–2019," IDC, Market Forecast, Doc # 259205, Oct. 2015.

[13] M. A. B. Ahmad, Mining Health Data for Breast Cancer Diagnosis Using Machine Learning, PHD Thesis, University of Canberra, Australia, Dec. 2013.

[14] W. Raghupathi and V. Raghupathi, "Big Data Analytics in Healthcare: Promise and Potential," Heal. Inf. Sci. Syst., vol. 2, no. 1, p. 3, 2014.

[15] Transparency Market Research, "Electronic Health Records Solution Market (Web Based, Client Server Based, Software as Services) for Applications in Hospitals, Physicians Office, Ambulatory Centers - Global Industry Analysis, Size, Share, Growth, Trends, and Forecast 2015 - 2023," MarketersMedia, USA, 2016.

[16] P. Zikopoulos and C. Eaton, Understanding big data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2011.

[17] A. Gandomi and M. Haider, "Beyond the Hype: Big Data Concepts, Methods and Analytics," Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015.

[18] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani and S. U. Khan, "The Rise of 'Big Data' on Cloud Computing: Review and Open Research Issues," Inf. Syst., vol. 47, pp. 98–115, 2015.

[19] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen and S. Iyengar, "Computational Health Informatics in the Big Data Age: A Survey," ACM Comput. Surv., vol. 49, no. 1, pp. 1–36, 2016.

[20] B. Feldman, E. M. Martin and T. Skotnes, "Big Data in Healthcare - Hype and Hope," Dr. Bonnie 360 Degree (Bus. Dev. Digit. Heal., vol. 2013, no. 1, pp. 122–125, 2012.

[21] Oracle, "Oracle Big Data Products," [Online], Available at: https://www.oracle.com/big-data/products.html.

[22] Hadoop, "Apache Hadoop," [Online], Available at: https://wiki.apache.org/hadoop.

[23] D. Miner and A. Shook, MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems, O'Reilly Media, Inc., 2012.

[24] N. Yuhanna and T. F. Wave™, "Market Overview: Big Data Integration," Forrester

Research, Inc. Reproduction Prohibited, 2014.

[25] A. Trnka, "Big Data Analysis," Eur. J. Sci. Theol., vol. 10, no. 1, pp. 143–148, 2014.

[26] J. Hurwitz, A. Nugent, F. Halper and M. Kaufman, Big data for dummies, John Wiley & Sons, 2013.

[27] Osnabrück University and IBM WATSON, "Flu-prediction Project," 2016, [Online], Available at: http://www.flu-prediction.com.

[28] Google, "Google Flu Trends," 2016, [Online], Available at: http://www.google.org/flutrends/about/

[29] D. Lazer and R. Kennedy, "What We Can Learn From The Epic Failure of Google Flu Trends," in: Wired, 2015.

[30] S. Yang, M. Santillana, and S. C. Kou, “Accurate estimation of influenza epidemics using Google search data via ARGO,” Proc. Natl. Acad. Sci., vol. 112, no. 47, pp. 14473–14478, Nov. 2015.

[31] B. Mole, "New Flu Tracker Uses Google Search Data Better Than Google," Scientific Method, USA, 2015.

[32] D. Lazer, R. Kennedy, G. King and A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," Science (80-.)., vol. 343, no. 6176, pp. 1203–1205, 14 March 2014.

[33] Builtinla, "Significant Benefits of Big Data Analytics In Healthcare Industry," Builtinla, 2016, [Online], Availavle at: http://www.builtinla.com/blog/significant-benefits-big-data-analytics-healthcare-industry.

[34] WHO, "Report of the Review Committee on the Functioning of the International Health Regulations (2005) in relation to Pandemic (H1N1) 2009," World Health Orgnaization, 2011.

[35] T. White, Hadoop: The definitive guide, O’Reilly Media, Inc., 2012.

[36] R. C. Taylor, "An Overview of the Hadoop/MapReduce/HBase Framework and Its Current Applications in Bioinformatics," BMC Bioinformatics, vol. 11, no. Suppl. 12, p. S1, 2010.

[37] The Apache Software Foundation, "Apache Pig!," 2016, [Online], Available at: https://pig.apache.org/. [Accessed: 05-Aug-2016].

[38] The Apache Software Foundation, "Apache Hive," 2016, [Online], Available at: https://hive.apache.org/. [Accessed: 05-Aug.-2016].

[39] "Apache Sqoop - Overview : Apache Sqoop," [Online], Available at: https://blogs.apache.org/sqoop/entry/apache_sqoop_overview. [Accessed: 24-Dec.-2016].

[40] "Introducing Sqoop - Cloudera Engineering Blog," [Online], Available at: http://blog.cloudera.com/blog/2009/06/introducing-sqoop/, Accessed: 24-Dec.-2016.

[41] A. Gupta, Learning Apache Mahout Classification, Packt Publishing Ltd., 2015.

[42] "5 Healthcare applications of Hadoop and Big data," DeZyre, 16 March 2015.

[43] K. J. Cios and G. William Moore, "Uniqueness of Medical Data Mining," Artif. Intell. Med., vol. 26, no. 1, pp. 1–24, 2002.

[44] "Methods for De-identification of PHI | HHS.gov," Health Information Privacy, [Online], Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-dentification/index.html#guidancedetermination, Accessed: 16 Jan. 2017.

[45] S. White, "A Review of Big Data in Health Care: Challenges and Opportunities," Open Access Bioinformatics, vol. 6, pp. 13–18, 2014.

[46] O. Tene and J. Polonetsky, "Privacy in The Age of Big Data: A Time for Big Decisions,"

Stanford Law Review, 2012, [Online], Available at: https://www.stanfordlawreview.org /online/privacy-paradox-privacy-and-big-data/.

[47] S. Kaisler, F. Armour, J. A. Espinosa and W. Money, "Big Data: Issues and Challenges Moving Forward," The 46th Hawaii International Conference on System Sciences (HICSS), pp. 995–1004, 2013.

[48] S. Kaisler, W. H. Money and S. J. Cohen, "A Decision Framework for Cloud Computing," The 45th Hawaii International Conference on System Sciences, pp. 1553–1562, 2012.

[49] J. Moon *et al.*, "Clinical Data Linkages in Spinal Cord Injuries (SCI) in Australia:," in: Big Data Analytics in Bioinformatics and Healthcare, vol. 43, no. 10, IGI Global, 1AD, pp. 392–405.

[50] A. W. Toga and I. D. Dinov, "Sharing Big Biomedical Data," J. Big Data, vol. 2, no. 1, p. 7, Dec. 2015.

[51] P. Moghe, "6 Hidden Challenges of Using the Cloud for Big Data and How to Overcome Them," Insider, 2016, [Online], Available at: http://thenextweb.com/insider/2016/04/12/6-challenges-cloud-overcome/.

[52] S. Robinson, "The Storage and Transfer Challenges of Big Data," MIT Sloan Management Review, 2012, [Online], Available at: http://sloanreview.mit.edu/article/the-storage-and-transfer-challenges-of-big-data/. [Accessed: 16-Jan-2017].

[53] S. Debortoli, O. Müller and J. vom Brocke, "Comparing Business Intelligence and Big Data Skills," Bus. Inf. Syst. Eng., vol. 6, no. 5, pp. 289–300, Oct. 2014.

ملخص البحث:

يتولّد في كوكبنا حجم هائل من البيانات في شتى المجالات، مثل: وسائل التواصل الاجتماعي، والصناعات، وأسواق المال، وأنظمة الرعاية الصحية. ويمكن لهذا الكم الهائل من البيانات أن يجلب الفوائد والمعرفة للأفراد والحكومات والصناعات وأن يساعد في اتخاذ القرارات. في مجال الرعاية الصحية، يتم توليد حجم هائل من البيانات من مقدمي الرعاية الصحية وتخزينها في أنظمة رقمية. وبذلك تكون البيانات أكثر قابلية للوصول من أجل الرجوع إليها أو استخدامها مستقبلاً. وتتلخص الرؤية النهائية لاستخدام البيانات الضخمة في مجال الرعاية الصحية في دعم عملية تحسين الخدمة المقدمة من موفّري الرعاية الصحية، والتقليل من الأخطاء الطبية، وتقديم المشورة عند الحاجة. تقدم هذه الورقة مراجعة نقدية لبعض تطبيقات البيانات الضخمة في الرعاية الصحية؛ مثل مشروع توقُّع الزكام الذي تبنّاه معهد العلوم الإدراكية. من جهةٍ أخرى، تصف هذه الورقة قيمة البيانات الضخمة في مجال الرعاية الصحية وتحللها وتعكسها. وفيها تم تعريف البيانات الضخمة بناءً على المصطلحات المتفق عليها لدى الباحثين والمهتمين في هذا المجال. كذلك، تبين هذه الورقة التوقعات المتعلقة بالعائد الاقتصادي للبيانات الضخمة لعام 2017، إلى جانب عائداتها التاريخية في ثلاثة مجالات أساسية هي: الخدمات، والمعدّات، والبرمجيات. وقد تم استعراض دورة إدارة البيانات الضخمة ومناقشة الجوانب الأساسية للبيانات الضخمة في مجال الرعاية الصحية (الحجم، والسرعة، والتنوع، والصدق). وفي الختام، تم إجراء مناقشة لبعض التحديات التي تواجه الأفراد والمنظمات في عملية الاستفادة من البيانات الضخمة في الرعاية الصحية؛ مثل: ملكية البيانات، وخصوصيتها وأمنها، والربط السريري للبيانات، والمسائل المتعلقة بتخزين البيانات ومعالجتها، والمتطلبات التي يجب توافرها بشأن مهارات العاملين في هذا المجال.