


A Brief Survey on Big Data in Healthcare

Ebru Aydindag Bayrak, Istanbul University-Cerrahpaşa, Turkey

 <https://orcid.org/0000-0002-2637-9245>

Pinar Kirci, Bursa Uludağ University, Turkey

ABSTRACT

This article presents a brief introduction to big data and big data analytics and also their roles in the healthcare system. A definite range of scientific researches about big data analytics in the healthcare system have been reviewed. The definition of big data, the components of big data, medical big data sources, used big data technologies in present, and big data analytics in healthcare have been examined under the different titles. Also, the historical development process of big data analytics has been mentioned. As a known big data analytics technology, Apache Hadoop technology and its core components with tools have been explained briefly. Moreover, a glance of some of the big data analytics tools or platforms apart from Hadoop eco-system were given. The main goal is to help researchers or specialists with giving an opinion about the rising importance of used big data analytics in healthcare systems.

KEYWORDS

Big Data, Big Data Analytics, Healthcare System, Medical Big Data

DOI: 10.4018/IJBDAH.2020010101

This article, originally published under IGI Global's copyright on April 17, 2020 will proceed with publication as an Open Access article starting on January 18, 2021 in the gold Open Access journal, International Journal of Big Data and Analytics in Healthcare (converted to gold Open Access January 1, 2021), and will be distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

1. INTRODUCTION

The technological developments helped us in producing more data that cannot be easily processed with currently available technologies. Thus, a new term 'big data' is created to describe the data that is large and not processed. Healthcare systems are generating huge amounts of data that present many positive and negative situations at the same time. For this reason, big data management and its analysis in healthcare sector are important (Dash et al., 2019).

Healthcare data increase day by day with the improvement of technology. The correct analysis of this data will increase the quality of maintenance and reduce the costs. This kind of data (big data) have some features such as high volume, variety, high speed production etc. Because of these features, analyzing data with traditional hardware and software platforms are pretty hard. Hence, choosing appropriate platform for analyzing and managing big data is very important (Nazari et al., 2019).

Considering the studies related to big data and big data analytics in the field of health in the literature, it is seen that quite a lot of studies have presented. Especially in recent years, it has been seen that there is a great increase in the number of studies on this subject. Some of them can be expressed as follows.

Galetsi et al. (2020) have studied on big data analytics in healthcare. Theoretical frameworks, techniques and prospects about big data analytics have been explained. They have aimed to present a systematic overview of the literature in order to show how much big data analytics has managed to contribute the healthcare system. Shilo et al. (2020) characterized health data by several axes that represent different components of the data. They described the potential and hardship of using big data in healthcare resources. They aimed to contribute to the continued argument of the potential of big data resources to improve the understanding of health and disease. McCall (2020) reviewed the interest in the use of big data in healthcare. Many big data analytics examples were explained for health in Silicon Valley. Alghunaim and Al-Baity (2019) studied on the problem of breast cancer prediction in the big data context. Support Vector Machine, Random Forest and Decision Tree used for machine learning classification by applying on each dataset. Apache Spark was used as a big data framework. Also, big data framework Spark was compared with WEKA traditional data processing environment. Bayrak and Kirci (2019) studied on intelligent big data analytics and machine learning systems for early diagnosis of neurological disorders. Many researches about intelligent big data analysis were reviewed in their study. Also, most used platforms or tools for big data analytics was explained. Carnimeo et al. (2019) was aimed to study a new health care network based on big data analytics for Parkinson's disease (PD). According to healthcare network, collected data during motor examinations of PD patients were analyzed and acquired knowledge was used to create a diagnostic report for patients. Dhayne et al. (2019) searched the topic of big medical data integration solutions. Data integration technologies, tools and applications was examined in the study. They focused on finding the strength and weakness of data integration technologies. Especially, they explained data integration

and it is the most important factor for healthcare sector. El Hanafi et al. (2019) studied on characterization of big data platforms for medical data. They described the big data environment with different components based on Hadoop Ecosystem. It was claimed that this system can be useful for helping the doctors to pursue their patients remotely. Nazari et al. (2019) investigated definition of the big data and the big data sources. In concern with big data, the advantages and challenges of big data, big data applications, big data analysis and big data platforms were explained. Palanisamy and Thirunavukarasu (2019) reviewed various healthcare frameworks and summarized their important ideas to learn the impact of big data in healthcare. The implications of big data tools in enhancing healthcare frameworks were considered. The big data tools were grouped data integration tools, machine learning tools, scalable searching and processing tools, visual data analytical tools and real-time and stream data processing tools. Uçar and İlkılıç (2019) worked on epistemological and ethical issues of the big data used in healthcare. Especially rather than whether big data was used in healthcare or not, they asked the question of “how” and in which conditions and in what moral lines should people use big data? Wang and Alexander (2019) explained big health data, big data in healthcare systems, applications, benefits and challenges of Big Data Analytics in healthcare system. They also presented a comparison of tools used for analyzing big data. Bahri et al. (2018) was aimed to explain big data technologies on the performance and outcomes of healthcare system. They explained big data process, technologies, and big data applications in healthcare sector. Big data application on healthcare was classified in five groups as Healthcare monitoring, healthcare prediction, recommendation systems, healthcare knowledge system and healthcare management system. Chiroma et al. (2018) surveyed the progress on Artificial Neural Networks (ANN) for big data analytics. They examined the application of ANN approaches on big data analytics. They explained that their study can be used by researchers as a criterion for future. Kouanou et al. (2018) studied on big data analytics for biomedical images and showed examples that were reported in the literature and new methods used in processing. They aimed to present a workflow for the management and analysis of biomedical image data which is based on big data technology. They designed two architectures to perform the image classification step. First architecture was based on the Hadoop framework and the second one was based on the Spark. Karabay and Ulaş (2017) described different types of tools that were mostly used in analysis of big data. The usage of big data, big data analyse methods and various big data technologies were mentioned in their study. Also, big data processing tools was explained and they were compared in terms of features such as operating system support, speed, real-time analysis and scalability etc.

The presented chapter is organized as follows. In Section 2, what is big data, the components of big data, big data sources and big data technologies in present are explained. In Section 3, big data analytics in healthcare system and some of the big data analytics tools /platforms are additionally presented. Section 4 continues with discussion and the chapter ends with conclusion by proposing concluding remarks in Section 5.

2. BIG DATA

IBM explains that every day 2.5 exabytes (EB) of data are created. CISCO estimates that, by 2020, 50 billion devices will be connected to internet and networks. The costs of enterprises on Information Technology (IT) infrastructure of the digital universe and telecommunications will increase in the rate of 40% between 2012 and 2020. And, Big Data will be responsible of nearly 40%. Furthermore, International Data Corporation (IDC) reckons that the 23 percent of the information in the digital world (or 643 EB) would be beneficial for Big Data. It comprises of data originated from embedded and medical devices, surveillance footage, entertainment, social media, as well as consumer images (Akoka et al., 2017).

The term ‘big data’ is defined in lots of works, and a globally adopted definition of big data has not been achieved yet in the research community. Big data is a subject which is growing with great popularity and can be described as a huge dataset that is characterized based on the ‘five V’s’ (Chiroma et al., 2018):

1. Volume means that the size of the dataset is very large;
2. Variety means that the data set is in different forms;
3. Velocity means that the content of the data adapts constantly;
4. Veracity means that the dataset has many options or interoperation variables in a mixed analysis;
5. Value means that the values in the dataset are huge and the density is very low (Figure 1).

Figure 1. The five components of big data (Adapted from Nazir et al. 2019)



Although the three components of the Big Data concept were met by the year it was created, today, new facts have added to these items and the number of components were increased to 5V. In fact, now it is considered as big data has 7V components rather than a 5V component (Çevik ve Özdemir, 2018):

6. Data Validity indicates that data sets may be valid for any application but not suitable for another applications;
7. Data Variability refers to the exclusion or destruction of data whose storage period has expired.

Globally, the use of big data in healthcare keep on increasing, Google's parent company, Alphabet, spent US\$2.1 billion to acquire FitBit. All over the world, healthcare system is interested in the use of big data and Silicon Valley is the one important part of this fact. Fifteen studies used big data in Silicon Valley, they are summarized in the following (McCall, 2020):

1. Komodo Health, traced the journey of 320 million de-identified US patients to understand health and disease at a scale;
2. Google Verily is a subsidiary company of Google's Alphabet and it purposes to make the world's health data beneficial, thus people will be happy and have healthy lives;
3. Helix uses its genome-sequencing possibilities to map sickness progression and identify new interference for sickness;
4. Ellipsis Health improved a vital sign tool for mental health and wellness for the detection of anxiety and depression;
5. Catalia Health system has a Wellness Coach robot which is called as Mabu. It holds daily and autonomous data that is derived from artificial intelligence generated conversations from patient;
6. Human Dx is both an educational tool and an AI-based decision tool that has big data obtained from clinicians and trainees globally;
7. Flatiron Health purposes to learn from the experience of every patient with cancer in its network by obtaining data from electronic health records;
8. PryAmes is a sensor platform for continuous blood pressure monitoring;
9. LunaDNA is a community-owned and health data sharing platform for health research;
10. Evidation is a virtual research site that collects big data from people's mobile apps, wearable sensors and devices;
11. Propeller Health is a digital health platform company for searching chronic obstructive pulmonary and asthma diseases;
12. Verana Health gathers the largest clinical databases to contribute researches;
13. Tidepool is an organization that has data of diabetes;
14. Bigfoot Biomedical is another start-up in the diabetes field;
15. Freenome uses next generation blood tests based on artificial intelligence for early detection of cancer disease.

Several big data sources in healthcare are available, such as clinical data, electronic health records, biometric data, registration data, patient reports, internet data, image data, biomarker data and administrative data (Nazir et al., 2019). Also, healthcare stakeholders and big data sources are patients, medical practitioners, healthcare insurers, hospital operators, the studies about clinical and pharma (Palamisamy and Thirunavukarasu, 2019).

According to “Data Never Sleeps 7.0” project that was prepared by Domo (2019), by 2020 there will be 40x more bytes of data than stars in the observable universe. It can be found out with how much data is generated in every minute of every day with some of the most popular platforms and companies in 2019 (Figure 2). For example; YouTube users watch 4,500,000 videos, Instagram users post 55,140 photos, Twitter users send 511,200 tweets. When the global internet population growth is examined, internet population is increasing enormously year by year. As of January 2019, used percentage of internet reached 56.1% of the world’s population and this rate is equal to 4.13 billion people (Domo, 2019).

According to all of these presented information, it can be said that big data was formed in the early 2000s as a concept. Surely one of the most significant factors effecting the concept was presented at the Davos Summit in 2012, and “data” was stated as an economic value. Big data is a concept used for expressing the volumetric size of the data at first, now it gained another meaning with expressing all of the processes from the storage of data to the information. The concept of “big data analytics” is also frequently used instead of the big data concept (Uçar and İlkılıç, 2019).

Figure 2. The spectacular picture of big data increase. 7th edition of Data Never Sleeps is related to how much data is being created in every minute in 2019 [Adapted from Domo].



3. BIG DATA ANALYTICS IN HEALTHCARE

The research communities in computer science and statistics have many working fields but statistical computing and machine learning began to play a major role in data mining, because the origins of big data analytics can be traced back to the 1970s or before.

In the present day, the volume and scale of data have increased dramatically due to the risen capability of computing power and automation. The data are referred to as very large databases or massive data sets among the computer science and statistics. According to the period of appearance of several methodologies about data analysis: data mining, knowledge discovery in databases (KDD) and statistical learning, this part is the first wave of big data analytics (Figure 3). Since the year of 2000, big data analytics have been successfully adopted in many disciplines: business schools, management schools, and informatics (bioinformatics, health informatics, systems informatics etc.). This period is called as the second wave (Tsui et al., 2019).

Big data analytics is the use of enhanced analytic techniques against huge, various data sets that include structured, semi-structured and unstructured data, from several sources, and between size of terabytes to zettabytes. Big data is a term applied to data sets whose size or type overcome the ability of classic relational databases to catch, run and handle the data with low latency. Analysis of big data give a chance to researchers and business users for having better and faster decisions by using data that was inaccessible or unusable data before (IBM,2020). Big data analytics refers to the use of a cutting-edge analytical tools to be able to effectively analyze and get information from big data that is high in the way of variety, volume, and velocity (Ghasemaghahi, 2020).

The analytical abilities of big data techniques and technologies that can be acquired from stored big data are useful for medical diagnosis, making predictions, resource allocations, recommendations and personal treatment plans. Big data process can be grouped as four steps: big data generation, big data acquisition, big data storage and big data analysis (Figure 4). This process is also called as big data chain value. (Bahri

Figure 3. The historical development process of big data analytics [Adapted from Tsui et al., 2019]

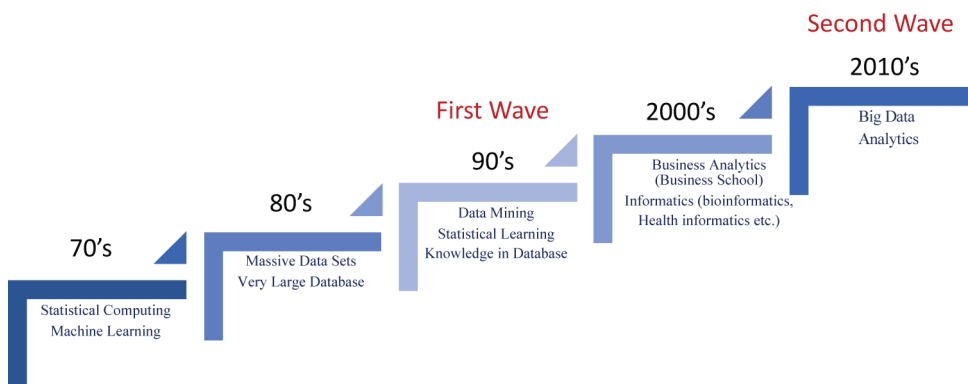
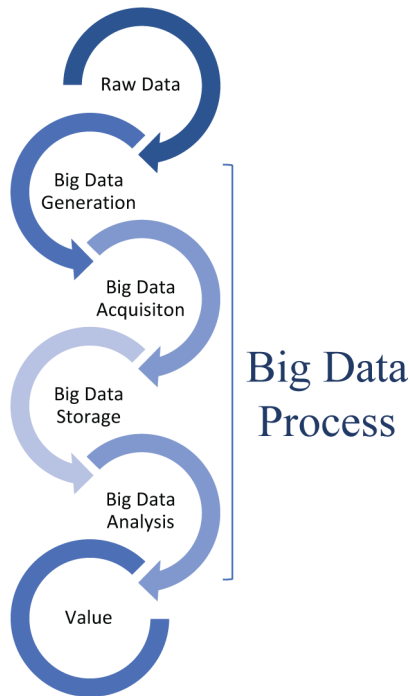


Figure 4. Big data process can be grouped in 4 steps that can be called as Big data chain value [Adapted from Bahri et al., 2018]



et al. 2018). If the data belong to healthcare industry, then the intensive and interactive dynamic big data platforms can be used, such as innovative technologies and tools to improve patient care and services in healthcare (Galetsi et al., 2020).

The most challenging parts for big data in healthcare are data privacy, data leakage, data security, effective use of big medical data, information security, wrong use of health data and misunderstanding unstructured clinical notes etc. Big data have a great opportunity to improve healthcare management and increase healthcare industry to a higher level (Bhattacharya, 2018).

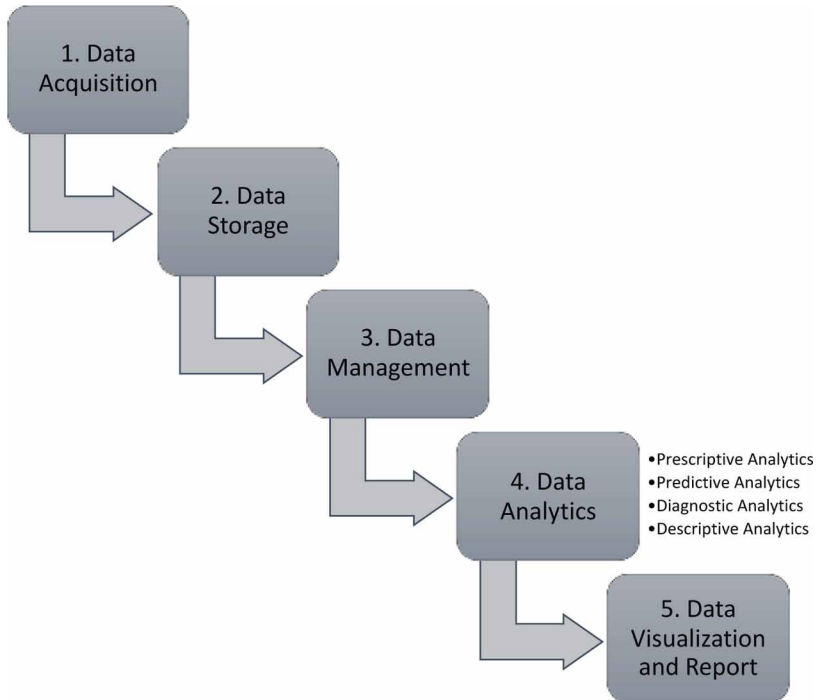
Big data analytics in healthcare system have been grouped as five processes: Data Acquisition, Data Storage, Data Management, Data Analytics and Data Visualization & Report. Also, the Figure 5 shows the process of big data analytics in healthcare management (Senthilkumar et al., 2018).

The big data analytics in healthcare system is in the early stage of its evaluation. For this reason, some handicaps and problems may emerge in the application areas. Any development about big data analytics can be much better for some important issues such as; accurate classification of a disease, quick decision making for a sickness and early diagnosis of many health problems.

3.1. HADOOP

Apache Hadoop is a well-recognized Big Data technology that has been used by an important supporting community. It was proposed to prevent the complexity and the

Figure 5. Big Data Analytics process in healthcare system [Adapted from Senthilkumar et al., 2018]



low performance encounter when Big Data is processed and analyzed with using traditional technologies (Oussous et al., 2018).

Among the software or platforms used in big data analytics, Hadoop software architecture, primarily uses Map-Reduce technology and HDFS (Hadoop Distributed File System). Hadoop is currently used in big data analytics for many leading IT companies, including Amazon, Yahoo, Facebook, Twitter, LinkedIn, IBM and Adobe. Hadoop architecture has two basic components, these are HDFS and Map-Reduce. The operations that will be performed in Hadoop are typically written as a job and given to the HDFS server cluster. HDFS works based on the cluster computing. In the cluster computing, it also means parallel processing architecture, servers with nodes are kept in rack cabinets (Aktan, 2018).

Hadoop is a software framework and it is developed to handle certain types of big data sets on a distributed system. It is developed with the inspiration of Google MapReduce and Google file system to process big-scale data between clustered computers. Hadoop is an open source ecosystem that combines a distributed file system called Hadoop Distributed File System (HDFS) with Hadoop MapReduce features (Karabay and Ulaş, 2017).

The core components of Hadoop HDFS, YARN and MapReduce can be explained as following.

3.1.1. HDFS (Hadoop Distributed File System)

HDFS is the file system component of Hadoop eco-system which stores file system metadata and it is known as blocks. HDFS is designed as a cost-effective and fault tolerant structure. HDFS gives file permission and authentication for all files. HDFS comprises of the name node and data nodes. HDFS works based on a master-slave architecture. A HDFS cluster contains a single name node, which is a master server that conducts the file system namespace and directories in the form of hierarchy. Data node contains two files. The first file is made up of data and the second file comprises of block's generation stamp (Hussain et al., 2019).

3.1.2. YARN (Yet Another Resource Negotiator)

YARN has been projected as the resource management framework of Hadoop. Hadoop Yarn ensures effective management of the resources. The purpose of the resource management technology is to partition system resources to various applications running in a Hadoop cluster. It is also used to schedule the conduct of tasks on different parts of clusters (El Hanafi et al., 2019). YARN is the central resource that was used by Hadoop. It traces the cluster nodes and all relevant processing operations. YARN permits Hadoop to perform operational activities without necessity of batch tasks to finish. Also, it has four fundamental components that are Resource Manager, Node Manager, Application Master and Container (Baig et al., 2019).

3.1.3. MAP REDUCE

MapReduce was suggested by Google for the first time. The map function filters the data to be processed and the Reduce function returns the analysis of the data processing as a result. When they get together, MapReduce system was appeared (Karabay and Ulaş, 2017). For processing big data, MapReduce is a well-known method to perform distributed storage and parallel computing. MapReduce is a programming model and an associated implementation for generating and processing big datasets and is convenient for semi structured or unstructured data (Dhayne et al., 2019).

The MapReduce, is used to process large data files in HDFS, it uses two functions for processing data. Map function is used to analyze and filter the data, and the Reduce function is used to get results from the data (Demirci, 2018).

4. SOME APACHE HADOOP TOOLS

4.1. Apache Spark

Apache Spark is a big data frame for the quick processes of datasets on various workloads. It can deal with interactive, batch, iterative and streaming data. Spark is accepted as 100 times faster in memory and 10 times faster on disk than Hadoop which is one of the most famous big data environments (Alghunaim et al., 2019). Apache Spark stores and processes data in memory. Spark not only processes data, but also provides Spark Machine Learning (MLlib) for machine learning, GraphX for

graphic processing, Spark SQL components for SQL support and Spark Streaming for streaming data (Figure 6). It also offers the possibility for developing with using languages such as R, Scala Java, Python. Thus, they make the application provide great convenience for development (Demirci, 2018).

4.2. Apache Storm

Storm is an open source framework to analyze big data in real time and it is consisted of spouts and bolts. In a storm program, a combination of a spout and a bolt is called topology. Spout can produce data or install data from an input and Bolt processes input streams and generates output streams (Benhlima, 2018). Apache Storm is another option for using the substitute of Hadoop MapReduce when we need real-time and heavy processing. Thereby, when our analytic solution is expected to process huge data as fastly, Apache Storm is the best performer for presenting such solutions (Harerimana et al., 2018).

4.3. Apache Cassandra

Cassandra is an Apache project, it is built on a distributed database system. It has NoSQL system and it is developed by Facebook for inbox searching. It has two million columns in a single row. It is highly scalable, accessible, robust and reliable and has zero point of failure. Cassandra is used by several companies such as Facebook, Twitter, and Amazon. It is inapplicable, if the use of data aggregation operation, sub query, and join operations are needed (Baig et al., 2019).

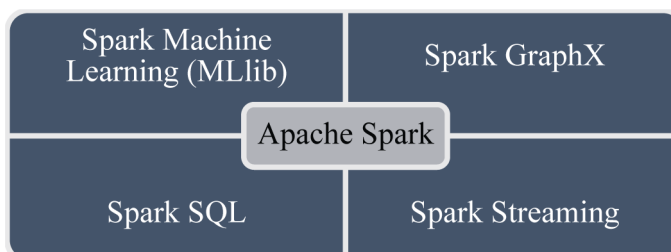
4.4. Apache Flink

Flink is a framework for all components of Hadoop eco-system and it is the framework for Streaming data. It has many advantages such as processing of data without latency and solving the memory exception problem. Flink also interacts with lots of devices that have several storage systems for processing the data, and it also optimizes the program before practice (Kumar et al., 2019).

4.5. Apache Zookeeper

Zookeeper can be used by different applications to coordinate the distributed processing of Hadoop clusters. It maintains the common objects needed in large cluster

Figure 6. The ecosystem of Apache Spark [Adapted from Spark, n.d.]



environments, including configuration information and the hierarchical naming space. Zookeeper also provides the reliability of application. If an application master dies, Zookeeper gets involved to create a new application master to be able to keep on the tasks (Kumar and Singh, 2018).

4.6. Apache Hive

Apache Hive is a data warehouse system, it is formed by Facebook in order to ease the usage of Hadoop.

It has an interactive interface with a variety of functions that are useful for data analysis and also, it is mostly used for structured data. The collected data is stored in a structured database that is easy and understandable for all users. Apache Hive database is administrated with a HQL language that has the same syntax like SQL language. HQL transforms queries into MapReduce jobs that are processed like batch tasks (Bahri et al., 2018). Hive is a data warehouse framework, it is used for querying and data analysis on Hadoop eco-system and it is developed by Apache. It uses HiveQL that is similar to SQL, to manage and query structured data. Hive's most important property is its convenience about translating the written codes. The codes are written in Hive and they are converted into Java MapReduce codes at the background. Because of this property, there is no need to learn Java program (Karabay and Ulaş, 2017).

4.7. Apache Kafka

Apache Kafka is a distributed streaming platform that has three key capabilities. These are publishing and subscribing streams of records, store streams of records by a fault-tolerant way and processing streams of records when they happen (Apache Kafka, n.d.). Apache Kafka is the storage platform for non-structured data. This platform is frequently used in real-time data processing and it can adjust how many days of data will be kept in the topic with based on the given configurations. Owing to this configuration, the data thrown into Kafka environment will not occupy a nonessential space in the system after processing (Özer, 2019).

4.8. Apache Pig

Apache Pig is a large-scale data analysis platform based on Hadoop eco-system. It ensures a language called PigLatin that has similar features as SQL language. The compiler of PigLatin turns SQL-like data analysis requests into a series of optimized MapReduce transactions. Apache Pig offers easy operations and programming interfaces for massive, complex data-parallel computing. For Apache Pig's main structure and running process it can be said that it just like Hive (Yu and Zhou, 2019).

4.9. Mahout

Mahout is an open source machine learning library which is developed by Apache. Mahout can be added on top of Hadoop for performing algorithms by way of MapReduce. Also, it is designed to study on their platforms. It has the advantage of providing scalable and influent implementation of large scale machine learning

applications and algorithms. Mahout library ensures to use analytical capabilities and multiple optimized algorithms. For instance, it offers libraries for classification, clustering, frequent pattern mining, collaborative filtering and text mining. Besides, additional tools involve dimensionality reduction, topic modeling, text vectorization, similarity measures, a math library and more than this (Oussous et al. 2018).

5. OTHER BIG DATA ANALYTICS TOOLS /PLATFORMS

SAP HANA was discovered by SAP company to process big data in real time. It provides a fast operation by using the in-memory database technology that contains continuously increasing big data. Owing to its in-memory technology, it has the ability to make big, instant, comparative, fast decisions and to analyze in real time. *SAP HANA* is not just a software that performs data control, analysis and integration processes, also with hardware it is a bundled data platform (Karabay and Ulaş, 2019).

BigML, is a scalable and programmable machine learning platform that ensures several tools to perform machine learning tasks such as classification, clustering, regression, association rules and anomaly detection. It can integrate the property of machine learning with cloud infrastructure to build cost-effective applications with high flexibility, reliability and scalability (Palanisamy and Thirunavukarasu, 2019).

KNIME is called as Konstanz Information Miner, it is an open source tool for data analytics, integration and reporting platform. It integrates different elements for data mining and machine learning through its modular data pipelining concept. Its graphical user interface allows the assembly of nodes for data pre-processing, data modelling, data visualization and data analysis. Since 2006, it has been widely used in pharmaceutical industry researches, but now it is also used in different areas like healthcare systems, customer data analysis, financial data analysis and business intelligence systems (Sampathrajan,2018).

6. DISCUSSION

In this chapter, it is aimed to present a specific information to the researchers by examining the studies on big data analysis in the field of healthcare. The application of big data analytics in health were reviewed. The most striking of these studies are tried to be summarized. It has been observed that the studies of big data analytics have been conducted on many different health problems (cancer, neurological diseases, etc.). How big data is defined in various sources is given and also big data studies in the field of health in Silicon Valley are mentioned.

The stages of big data analysis and big data analytics applied in the field of health are specified. In addition, various big data analytics platforms are mentioned. The number of presented studies, the utilized tools and platforms that can be used for big data analytics is quite high. This chapter also focuses on Hadoop architecture. Some components of the Hadoop eco-system have been studied under separate titles. It is clearly seen that big data has grown and continues growing with an incredible rate

until today. Therefore, it can be said that the studies in big data and big data analytics will continue to increase.

7. CONCLUSION

This chapter presents a brief approach on big data and big data analytics and also about their roles in healthcare system. A definite range of scientific researches about big data analytics in healthcare system is mentioned.

Apache Hadoop eco-system has been tried to be explained because it is mostly used as big data analytics for lots of studies. Its core components HDFS, YARN and MapReduce are explained and also some Hadoop tools used as big data analytics such as Spark, Storm, Mahout, Pig, Hive etc. are examined. Moreover, a glance is taken over some of the big data analytics tools or platforms apart from Hadoop eco-system.

This chapter provides an opinion of how health (medical) big data can be used as a source for big data analytics. The main aim behind this chapter is to assist the researchers or experts to provide an idea and understand about big data analytics which can be used for solving several health problems.

The big data analytics in healthcare system can be said that is in the early stage of its evaluation. For this reason, some handicaps and problems can happen in the field of application areas. Each one of the presented developments about big data analytics can provide much better ideas for some important issues such as; accurate classification of a disease, quick decision making for a sickness and early diagnosis of different health problems.

REFERENCES

- Akoka, J., Comyn-Wattiau, I., & Laoufi, N. (2017). Research on Big Data—A systematic mapping study. *Computer Standards & Interfaces*, 54, 105–115. doi:10.1016/j.csi.2017.01.004
- Aktan, E. (2018). Büyük veri: Uygulama alanları, analitiği ve güvenlik boyutu. *Bilgi Yönetimi*, 1(1), 1–22. doi:10.33721/by.403010
- Alghunaim, S., & Al-Baity, H. H. (2019). On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context. *IEEE Access: Practical Innovations, Open Solutions*, 7, 91535–91546. doi:10.1109/ACCESS.2019.2927080
- Apache Kafka. (n.d.). <https://kafka.apache.org/intro.html>
- Apache Spark. (n.d.). <https://spark.apache.org/>
- Bahri, S., Zoghlami, N., Abed, M., & Tavares, J. M. R. (2018). Big data for healthcare: A survey. *IEEE Access: Practical Innovations, Open Solutions*, 7, 7397–7408. doi:10.1109/ACCESS.2018.2889180
- Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2019). Big Data Tools: Advantages and Disadvantages. *Journal of Soft Computing and Decision Support Systems*, 6(6), 14–20.
- Bayrak, E. A., & Kirci, P. (2019). Intelligent Big Data Analytics in Health. In *Early Detection of Neurological Disorders Using Machine Learning Systems* (pp. 252–291). IGI Global. doi:10.4018/978-1-5225-8567-1.ch014
- Benhlime, L. (2018). Big data management for healthcare systems: Architecture, requirements, and implementation. *Advances in Bioinformatics*. PMID:30034468
- Bhattacharya, D. (2018). Big Data Management and Growth Enhancement. *International Research Journal of Engineering and Technology*, 5(10), 1769–1774.
- Carnimeo, L., Trotta, G. F., Brunetti, A., Cascarano, G. D., Buongiorno, D., Loconsole, C., & Bevilacqua, V. et al. (2019). Proposal of a health care network based on big data analytics for pds. *Journal of Engineering (Stevenage, England)*, 2019(6), 4603–4611. doi:10.1049/joe.2018.5142
- Çevik, Ö. Ü. N. K., & Özdemir, Ö. G. M. (2018, december). Büyük Veri: Tanımı ve Oluşumu. In *International Congress of Management Economy and Policy 2018 Spring Proceedings Book* (p. 29). Academic Press.
- Chiroma, H., Abdullahi, U. A., Alarood, A. A., Gabralla, L. A., Rana, N., Shuib, L., & Herawan, T. et al. (2018). Progress on Artificial Neural Networks for Big Data Analytics: A Survey. *IEEE Access: Practical Innovations, Open Solutions*, 7, 70535–70551. doi:10.1109/ACCESS.2018.2880694

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data*, 6(1), 54. doi:10.1186/s40537-019-0217-0

Demirci, C. (2018). *Büyük veri platformlarından hadoop ile örnek veri analizi*. Academic Press.

Dhayne, H., Haque, R., Kilany, R., & Taher, Y. (2019). In Search of Big Medical Data Integration Solutions-A Comprehensive Survey. *IEEE Access: Practical Innovations, Open Solutions*, 7, 91265–91290. doi:10.1109/ACCESS.2019.2927491

Domo. (n.d.). <https://www.domo.com/learn/data-never-sleeps-7>

El Hanafi, H. E. A., Afifi, N., & Belhadaoui, H. (2019). *Characterization of Big Data Platforms for Medical Data* (No. 1593). EasyChair.

Galetsis, P., Katsaliaki, K., & Kumar, S. (2020). Big data analytics in health sector: Theoretical framework, techniques and prospects. *International Journal of Information Management*, 50, 206–216. doi:10.1016/j.ijinfomgt.2019.05.003

Ghasemaghaei, M. (2020). The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage. *International Journal of Information Management*, 50, 395–404. doi:10.1016/j.ijinfomgt.2018.12.011

Hussain, T., Sanga, A., & Mongia, S. (2019). *Big Data Hadoop Tools and Technologies: A Review*. Available at SSRN 3462554

IBM. (n.d.). *Big data analytics*. <https://www.ibm.com/analytics/hadoop/big-data-analytics>

Karabay, B., & Ulaş, M. (2017). Büyük Veri İşlemede Yaygın Kullanılan Araçların Karşılaştırılması. *8th International Advanced Technologies Symposium (IATS'17)*.

Kouanou, A. T., Tchiotsop, D., Kengne, R., Zephirin, D. T., Armele, N. M. A., & Tchinda, R. (2018). An optimal big data workflow for biomedical image analysis. *Informatics in Medicine Unlocked*, 11, 68–74. doi:10.1016/j.imu.2018.05.001

Kumar, J. S., Raghavendra, B. K., & Raghavendra, S. (2019). Big data Processing Comparison using Pig and Hive. *International Journal on Computer Science and Engineering*, 7, 173–178.

Kumar, S., & Singh, M. (2018). Big data analytics for healthcare industry: Impact, applications, and tools. *Big Data Mining and Analytics*, 2(1), 48–57. doi:10.26599/BDMA.2018.9020031

McCall, B. (2020). 15 ways Silicon Valley is harnessing Big Data for health. *Nature Medicine*, 26(1), 7–10. doi:10.1038/s41591-019-0708-8 PMID:31932786

- Nazari, E., Shahriari, M. H., & Tabesh, H. (2019). BigData Analysis in Healthcare: Apache Hadoop, Apache spark and Apache Flink. *Frontiers in Health Informatics*, 8(1), 14. doi:10.30699/fhi.v8i1.180
- Nazir, S., Nawaz, M., Adnan, A., Shahzad, S., & Asadi, S. (2019). Big Data Features, Applications, and Analytics in Cardiology—A Systematic Literature Review. *IEEE Access: Practical Innovations, Open Solutions*, 7, 143742–143771. doi:10.1109/ACCESS.2019.2941898
- Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4), 431–448. doi:10.1016/j.jksuci.2017.06.001
- Özer, S. (2019). *Büyük Veri Teknolojileri ve Veri Madenciliği Yöntemleri ile Medikal Veri Analizi* (Master Thesis). Marmara University, Science and Engineering Institute.
- Palanisamy, V., & Thirunavukarasu, R. (2019). Implications of big data analytics in developing healthcare frameworks—A review. *Journal of King Saud University-Computer and Information Sciences*, 31(4), 415–425. doi:10.1016/j.jksuci.2017.12.007
- Sampathrajan, S. (2018). A Study of Big Data Practices in Various Open Source Tools. *International Journal of New Technologies in Science and Engineering*, 5(7), 27–34.
- Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: Challenges and promises of big data in healthcare. *Nature Medicine*, 26(1), 29–38. doi:10.1038/s41591-019-0727-5 PMID:31932803
- Tsui, K. L., Zhao, Y., & Wang, D. (2019). Big Data Opportunities: System Health Monitoring and Management. *IEEE Access: Practical Innovations, Open Solutions*, 7, 68853–68867. doi:10.1109/ACCESS.2019.2917891
- Uçar, A., & İlkılıç, İ. (2019). Büyük Verinin Sağlık Hizmetlerinde Kullanımında Epistemolojik ve Etik Sorunlar. *Sağlık Bilimlerinde İleri Araştırmalar Dergisi*, 2(2), 80–92.
- Wang, L., & Alexander, C. A. (2019). Big data analytics in healthcare systems. *International Journal of Mathematical. Engineering and Management Sciences*, 4(1), 17–26.
- Yu, J. H., & Zhou, Z. M. (2019). Components and Development in Big Data System: A Survey. *Journal of Electronic Science and Technology*, 17(1), 51–72.

Ebru Aydindag Bayrak is currently PhD student of Department of Engineering Sciences, Faculty of Engineering, the University of Istanbul Cerrahpaşa. She was received BSc degree from Department of Mathematics, Faculty of Science and Literature, Çanakkale Onsekiz Mart University at 2012 and also she was graduated pedagogical formation from Faculty of Education from there. Then, she received her MSc degree from Institute of Science, Istanbul University about Fractal Analysis of Active Fault Data in the San Andreas and the North Anatolian Fault Zones. Her poster presentation about Fractal Analysis of Active Fault Data in North Anatolian Fault Zone won first poster prize in meeting of Active Tectonic Research Group at 2015. Also she was graduated Department of Business Administration (BBA degree) from Istanbul University at 2016. In PhD studies she has working on machine learning techniques for the prediction of cancer. She continues to work on academic studies in conferences, journals and symposiums for both national and international.