First Libyan international Conference on Engineering Sciences & Applications (FLICESA_LA)
13 – 15 March 2023, Tripoli – Libya

Review of Big Data Analytics in Healthcare Sector

Mohamed ali whiba
College of IT
Dept. Mobile Computing,
University of Tripoli
Email : m.whiba@uot.edu.ly
Phone : 0925425016

Akram ali milad
College of Science
Dept. Computer Science,
University of Tripoli
Email : ak.milad@uot.edu.ly
Phone : 0925025434

Abstract: The term "big data" does not just mean big in size but also high in variety and velocity, making it hard to deal with it using traditional tools and techniques. Nowadays, (BDA) big data analytics is the most effective way to discover hidden patterns and knowledge from vast amounts of data and gain future insight to assist decision-makers in taking the right trend toward presenting better services. In general, Big Data Analytics (BDA) implementations in any domain are complex and resource intensive, with a high failure rate and insufficient guidance or strategy to assist practitioners in making sense of that massive amount of data. In this paper, we define big data in general, determine characteristics of big data for healthcare, and highlight important sources of big data in healthcare. In addition, recommend architecture and tools, identify the most common challenges facing BDA, and recommend a proposed system for BDA in the healthcare sector.

Keywords
Big Data, Healthcare, Big data analytics, Hadoop, Hadoop Spark.

I. 1.INTRODUCTION

Data sets expand quickly due to the regular collection of information by several information-sensing devices, including wireless sensor networks, mobile devices, aerial (remote sensing), software logs and records, cameras, microphones, and Radio-Frequency Identification (RFID) readers. In other words, big data is a field that describes how to evaluate data sets that are too huge or complicated for typical data processing systems to handle, how to gather information from them systematically, and how to deal with them.[1]

One of the most vital industries is the healthcare sector. It is also one of the biggest and fastest-growing sectors in the world. Although different electronic health records (EHRs) collect data in different ways—structured, unstructured, and semi-structured, the industry can produce and handle data at a startling rate. This type can be difficult when attempting to verify the truth or ensure the data's quality. The EHRs can offer a wealth of data that is available for analysis to advance our knowledge of disease causes and to deliver better, more individualized healthcare, but the data structures present a challenge to conventional methods of analysis. Therefore, adopting big data analytics technologies is necessary to transform the raw data into relevant and actionable information.[1]

II. BACKGROUND

2.1. Defining Big Data

"Big data" is a term used to describe massive amounts of data that are produced quickly and contain a lot of information. Unstructured sources of this data include streams of web clicks, social media platforms (Twitter, blogs, Facebook), call center call recordings, video recordings from stores, real-time information from various types of sensors, RFID, GPS, mobile phones, and other devices that identify and monitor something. Big Data is a potent digital data silo that is unstructured, acquired from a variety of sources, raw, and hard or even impossible to analyze using conventional techniques used so far for relational databases.[2]

2.2. Big Data Properties

The six "Vs," as shown in Figure 1, (value, volume, velocity, variety, veracity, and variability), are characteristics that define big data. The three of them (volume, velocity, and variety) are recognized as the main characteristics of big data. In addition to these six "Vs," some authors have established more than these six attributes as described below to explain the characteristics of big data (volume, velocity, and variety), which are recognized as the main characteristics of big data. Knowing how to measure big data requires an awareness of these characteristics. [1],[3]

Value
Variability
Volume
The 6 V's of big data
Veracity
Velocity
Variety

Figure 1: The 6 V'^{S} of big data.

- Volume (refers to the amount of data and is one of the biggest challenges in Big Data Analytics),
- Velocity (the speed with which new data is generated, the challenge is to be able to manage data effectively and in real-time),
- Variety (heterogeneity of data, many different types of healthcare data, the challenges to derive insights by looking at all available heterogeneous data in a holistically manner),
- Variability (inconsistency of data, the challenge is to correct the interpretation of data that can vary significantly depending on the context),
- Veracity (how trustworthy the data is, quality of the data),
- Visualization (ability to interpret data and resulting insights, challenging for Big Data due to its other features as described above).
- Value (the goal of Big Data Analytics is to discover the hidden knowledge from huge amounts of data).[2]

The open-source distributed data processing platform Apache Hadoop MapReduce, which is based on data-intensive computing and (No-SQL) Not only Structured Query Language data modeling approaches, is a very relevant and promising software platform for the development of applications that can manage big data in medicine and healthcare.[3]

2.3 Big data analytics

Big data analytics can uncover unknown patterns and trends, hidden correlations, and other important information from data (value). This theory is based on the data life cycle framework, which begins with data capture, continues through data transformation, and finishes with data consumption.[5]

BDA outcomes give analytics experts, predictive modelers, and data scientists the ability to analyze and look at massive amounts of data sets with a variety of data types that may be undiscovered using conventional analytics techniques. BDA may result in advantages in the marketplace, increased operational effectiveness, greater customer service, more fruitful new opportunities, and other benefits.[7]

III. Research Methodology

To accomplish the objectives of this study, we employed a quantitative strategy, more specifically, a content analysis of different cases, to classify and comprehend the capabilities of big data analytics and the potential advantages associated with their utilization. The next subsections discuss the methods used to collect the cases, the techniques used, and the steps involved in case analysis.

3.1 Cases collection

Our instances were derived from current and previous big data projects using data from a variety of sources, including case collections, print publications, practical journals, and reports from businesses, vendors, consultants, and analysts.

3.2 Source of the information

To locate relevant research publications, we looked at four databases: (1) IEEE Xplore, (2) ScienceDirect, (3) Springer, and (4) Scopus. The main search terms we used in these databases were "big data" or "big data analytics," as well as "healthcare" or "medicine".

3.3 Selection Criteria

The review method for this work is depicted in Figure 4. Five rules are followed: (i) search strategy, (ii) selection criteria, (iii) study selection procedure, (iv) quality assurance, and (v) qualitative and quantitative analysis. We concluded by outlining what was done in each phase.

Due to their open access and flexible date restriction policies, we chose Google Scholar and Science Direct as the main search engine platforms for gathering pertinent articles.

Only pertinent journal and conference papers, though, were downloaded.

Big data, big data analytics, big data in health care, Apache Spark, and Hadoop were the five key terms utilized in the search, according to the authors. However, using Google Trends, we see a number of inquiries that are relevant to our five keywords. The terms "big data and data analytics," "analytics of big data," "data analytics," "analytics," "Hadoop," "big data Hadoop," and "Hadoop spark" are only a few of those that were used in this study as auxiliary terms.

Based on an inclusion-exclusion criterion that we had agreed upon, the articles were evaluated. The following requirements must be met for inclusion: (i) the article must be written in English, (ii) it must deal with big data and BDA, (iii) it must have been published between 2015 and 2021, and (iv) it must have been published in a journal or conference. In contrast, the exclusion criteria were (i) papers published outside of a journal or conference and (ii) articles not published between 2015 and 2021. Even more, papers not published in Elsevier, Springer, or IEEE, review-based articles on BDA, Initially, in the primary screening phase, all the publications pertinent to the current study (big data and BDA) were carefully chosen. The checked articles (3030) were screened in Stage one according to our inclusion and exclusion criteria (see Fig. 2), we excluded inappropriate papers (i.e., papers not published in English and duplicate or overlapping publications). We then selected the remaining publications based on their title, abstract, publishers, and type of publishing, and we excluded any that had no relevance to the planned study. In the last round of screening, these articles were selected based on the abstracts using the Boolean AND operator on all of the defined search terms.

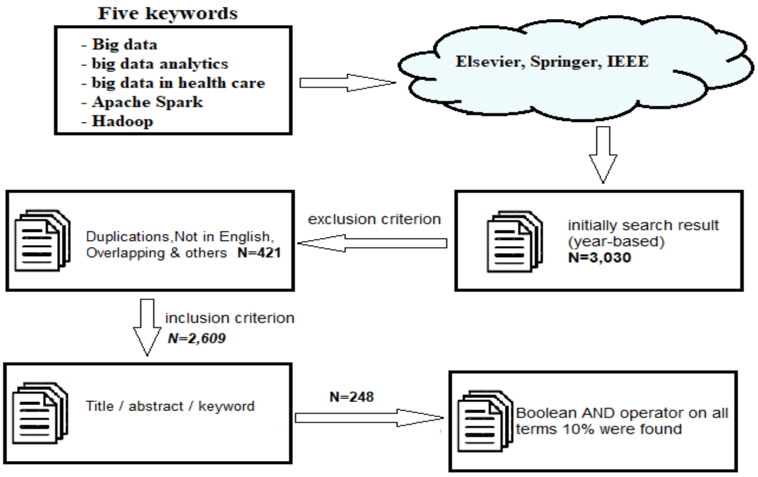
Five keywords
- Big data
- big data analytics
- big data in health care
- Apache Spark
- Hadoop
Elsevier, Springer, IEEE
initially search result (year-based) N=3,030
exclusion criterion
Duplications,Not in English, Overlapping & others N=421
inclusion criterion N=2,609
Title / abstract / keyword
N=248
Boolean AND operator on all terms 10% were found

Figure 2: Research Process.

IV. Result and Discussion

Based on this study, we found that many sources of big data are not processed by traditional data processing techniques, and these data are growing exponentially every moment. Therefore, the health care sector should pay attention to stockholders or decision makers to provide the best services in this sector by extracting valuable information from stacks of data generated from patients, sensors, electronic health records, patient records history, x-ray devices, etc.

This extraction is achieved through data analyst profession, which must have enough capabilities to analyze data using tools and platforms.

We will discuss the sources of big data and the tools and platforms that are most effective for extracting hidden patterns from big data. We will present the challenges faced by big data analytics.

4.1 Source of Health Care Big Data

The data acquired, gathered, and stored in the healthcare sector may be dispersed from diverse sources with varied structures, forms, and types. Data on physiological, behavioral, clinical, medical imaging, disease administration, prescription drug records, diet, and exercise factors are all included in healthcare big data. However, the majority of the studies under consideration agreed on the following common sources and types of big data streams in the healthcare sector [1]:

- Clinical data includes data from electronic medical records, hospital information systems, imaging centers, labs, pharmacies, and other businesses that provide healthcare services, as well as patient-generated health data, free-text notes from doctors, genomic data, and data from physiological monitoring.
- Biometric data is generated by various devices that monitor things like glucose levels, blood pressure, weight, etc.
- Financial data includes a complete accounting of all economic activity.
- Data from scientific research projects, i.e., results of studies on drugs, medical device design, and innovative treatment methods.
- Data provided by the patients, such as a description of their preferences and their degree of satisfaction, and data from systems for tracking their own activities, such as how much they exercise, sleep, eat, and so on.
- Data from social media.

As a result, there is promise for big data analysis, especially in the areas of raising medical care standards, saving lives, and reducing costs. By sorting through this maze of given association rules, patterns, and trends, healthcare service providers and other stakeholders will be able to offer patients more precise and insightful diagnoses, personalized treatment, patient monitoring, preventive medicine, support for medical research and the population's health, better quality medical services, and patient care while also being able to reduce costs[2],[7].

4.2 Big data analytics infrastructure and tools

Contrary to common belief, Hadoop is not the most effective or widely used technology in BDA, as shown by our findings. We found that Spark, followed by Hadoop is the DBA tool that researchers in this discipline use the most frequently. This result can be explained by the fact that Spark is more efficient and user-friendly for big data analytics than Hadoop MapReduce. Additionally, it is thought that Spark processes data more quickly than Hadoop [9],[10].

However, a comparison study reveals that Spark uses more memory during operation than Hadoop since it loads all processes into memory and caches them for a while [8],[9],[10]. As a result, this article advises basing your decision on these two platforms' various features. For example, cost, ease of use, memory restrictions, fault tolerance, performance level, data processing type, and security, and demonstrate their suitability for the project at hand and the organization's present and future demands. In conclusion, as more businesses and researchers have discovered the sweet spot to adopt these platforms, the open-source industry for the Spark and Apache frameworks has experienced tremendous growth. Table.1 shows comparison between Hadoop and Spark [11].

Table 1

HADOOP vs SPARK COMPARISON		
	HADOOP	**SPARK**
What is it?	Open-source framework for distributed data storage and processing	Open-source framework for in-memory distributed data processing and app development
Initial release	2006	2014
Supported Languages	Java	Scala, Java, Python, R
Processing methods	Batch processing, using hard discs to read/write data	Batch and micro-batch processing in RAM
Built-in capabilities	✓ File system (HDFS) ✓ Resource management (Yarn) ✓ Processing engine (MapReduce)	✓ Processing engine (Spark Core) ✓ Near real-time processing (Spark Streaming) ✓ Structured data processing (Spark SQL) ✓ Graph data management (GraphX) ✓ ML library (MLlib)
Best fit for	Delay-tolerant processing tasks, involving huge datasets	Almost instant processing of live data and quick analytics app development
Real-life use cases	✓ Enterprise archived data processing ✓ Sentiment analysis ✓ Predictive maintenance ✓ Log files analysis	✓ Fraud detection ✓ Telematics analytics ✓ User behavior analysis ✓ Near real-time recommender systems ✓ Stock market trends prediction ✓ Risk management

4.3 Spark Architecture

Data science regularly makes use of Apache Spark, an in-memory data processing engine with a functional flavor and superior semantics for iterative types of computation. Spark was developed to become the most efficient tool for data analytics and expanded to become a top-level Apache project with additional components such as Spark SQL, MLlib, and stream processing in addition to the fundamental processing capabilities. The RDD (resilient distributed dataset), which is a distributed sequence of objects with an implicit mechanism to accommodate failure, is the fundamental data construct used by Spark. To rebuild some of the missing data, Spark may repeat the operations on a subset of the data. The relational operators, like union, distinct, filter, and join, that are integrated into the Spark programming paradigm can be applied lazily to RDDs. Figure 3 shows the general Hadoop architecture with HDFS, Yarn and the processing engine Spark

.

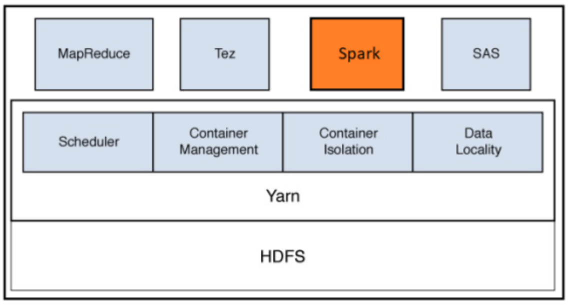

Figure 3 Hadoop architecture with HDFS, Yarn, processing engine Spark.

Apache Spark is a relatively new framework for distributed, in-memory data processing, as was already mentioned. Pre-processing of data is greatly facilitated by Spark's support for interactive data processing with support for Python or Scala.

The main abstraction of Spark is that of a resilient distributed dataset (RDD), which can be processed using any transformation logic written in Scala or Python as well as other relational algebra operators (such as select, filter, join, and group by).

A fresh, more practical abstraction called a DataFrame is built on top of this idea to make slicing and dicing data even simpler. This API is convenient to Python's Pandas library's practical slicing and dicing APIs.

As it continues to develop, Spark now offers a lot more capabilities than its original RDD-based processing engine[12] (as in Figure 4).

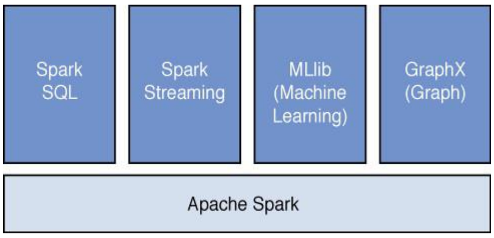

Figure 4 Spark architecture

Spark SQL provides an alternative to Hive for SQL on distributed datasets. The DataFrames API, which offers a more programmatic API for quickly and efficiently processing large datasets using relational algebra, is supported by Spark SQL, which is built on top of Spark Core. This API supports both conventional SQL queries and relational algebra. The ability to perform some processing in SQL and continue from there in Spark without writing the data to disk is one of the most interesting aspects of Spark SQL.

Spark Streaming, similar to Apache Storm, is a Spark component for creating scalable, fault-tolerant streaming systems. Slice-based data processing is one way this functionality is implemented, as is an extension to the standard Spark API. As opposed to the resilient distributed dataset (RDD) that the standard Spark dataset abstraction is, the discretized stream used by the Spark streaming dataset abstraction (DStream).On these discretized segments of the stream, processing can then take place.

Spark MLlib has an implementation of several machine learning methods across distributed datasets that is provided by the machine learning library, which is integrated with the Spark toolset. With each new Spark MLlib release, the library of supported algorithms expands, but many of the most popular algorithms, such as linear and logistic regression, support vector machines (SVM), decision trees and random forests, k-means clustering, singular value decomposition (SVD), and many others, are already present.

Spark GraphX offers a toolkit for graphs and graph-parallel computations on top of Spark, with support for popular algorithms like PageRank, label propagation, triangle count, and others.[12]

4.4 Big data analytics Challenges in Healthcare

With the rate of data explosion, the biggest difficulty in dealing with big data is that current or traditional systems are incapable of storing and processing data of this quantity and type. As a result, cloud computing can be used to tackle the storage problem, where small and medium-sized hospitals and care groups would be able to reduce cost and data storage difficulties even more, and they would be able to scale out storage, reduce infrastructure maintenance, and increase the availability of health care system services, such as those of cloud providers Amazon Web Services, Microsoft Azure, and Google Cloud Platform, which are the top cloud service providers that dominate the worldwide cloud market.

These three cloud providers have the experience and expertise to provide a dependable and feature-rich cloud platform. However, before committing to a certain cloud platform, you must conduct due diligence and analyze each platform to thoroughly grasp their capabilities and differences.[1],[13]

and other challenges in big data analytics in the healthcare sector and elsewhere Data ownership is a critical and ongoing issue in big data applications. Though petabytes of medical records normally belong to the healthcare providers, governmental healthcare systems, or hospitals that developed them, the information contained inside them does not. Patients, on the other

hand, believe that they own the data. This dispute may be resolved through the legal system unless healthcare practitioners have written permission from patients before using data for experiences or research aims.[14]
Furthermore, multiple studies [15, 16, 17] have highlighted the difficulty in dealing with this vast and unprecedented amount of data made up of various data types and even streaming data.
As a result, big data is high-dimensional, diversified, massive, complex, incomplete, amorphous, noisy, and erroneous, making data pre-processing in BDA difficult. However, it is essential to ensure that machine learning models perform well and with high accuracy.[6]

4.5 recommended system with big data analytics tools

The use of the cloud in conjunction with data has become a more recent trend in big data analytics technologies. A "big data in the cloud" option, such as software-as-a-service (SaaS), that provides an alluring alternative at a reduced price, has gained increasing traction with businesses. By 2016, using cloud computing services for big data analytics systems that enable real-time analytic capability and cost-effective storage will become a favored IT solution, according to Gartner's 2013 IT trend prediction.

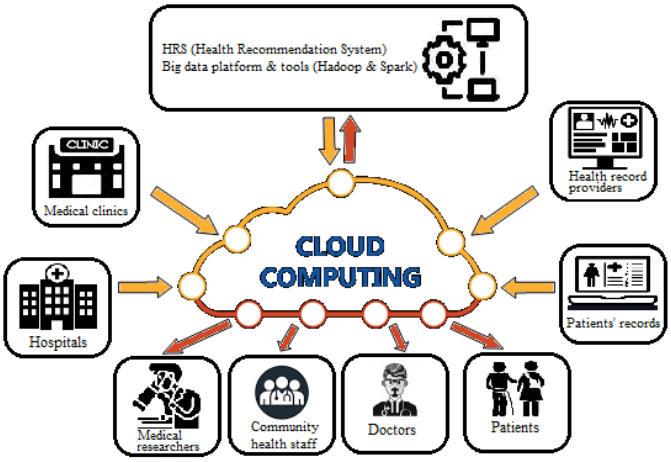

Figure 5 recommender system and its stakeholders

The primary trend in the healthcare sector is a switch from structured to semi-structured and unstructured data (e.g., home monitoring, telemedicine, sensor-based wireless devices, transcribed notes, images, and video).

The suggested recommender system for big data analytics is shown in Fig. 5, which is composed of eight stakeholders' components: (1) Hospitals, (2) Medical Clinics, (3) Patient Records, (4) Health Record Providers, (5) Medical Researchers, (6) Community Health Staff, (7) Doctors, and (8) Patients, in addition to two components of the recommender system for data analysis using Hadoop, Spark, and cloud computing for use in storing large amounts of data.
These components make up the big data analytics components that carry out specialized tasks, enabling healthcare administrators to grasp how to use big data implementations to transform healthcare data from numerous sources into useful clinical information.

All the data sources required to deliver the insights needed to support regular operations and resolve business challenges are included for the first four components. Structured data, like conventional electronic health records (EHRs), semi-structured data, like the logs of health monitoring devices, and unstructured data, like clinical photographs, are three categories of data. Depending on the format of the content, these clinical data are gathered from various internal or external sites and saved right away in cloud databases.[5]

In contrast, the recommender system performs analytics using Hadoop and Spark as tools for the data stored in cloud databases that contain databases of the type NO-SQL like Mongodb, and it transforms the data into useful information or clinical information to use it for the remaining four components (medical researchers, community health staff, doctors, and patients) for decision-making.

V. Conclusions

Our research has provided a deeper understanding of how healthcare firms may use big data analytics to transform data into useful information. For better illness treatment and diagnosis in medicine, big data's function is to build better prediction models with tools that can analyze and process massive amounts of data to get future insight.

The main limitation of this study is obtaining data from sources. To analyze big data, analysts must be able to collect, clean, and visualize row data in order to create models that can be used with the existing tools for analysis. Finally, we recommended a proposed system utilizing big data analytics and exploiting cloud computing to store and manage the vast amount of data and involve using tools and platforms such as Spark and Hadoop in the healthcare sector.

The reason for choosing the architecture Hadoop with Spark instead of Hadoop with MapReduce is that Spark is faster than MapReduce because it is based on memory for data processing. In contrast, MapReduce is based on disk for data processing.

References

[1] BIG DATA ANALYTICS IN HEALTH CARE: A REVIEW PAPER Maria Mohammad Yousef Department of Computer Science, Al-albayt University, Jordan ,Vol 13, No 2, April 2021.

[2] The use of Big Data Analytics in healthcare Kornelia Batko1* and Andrzej Ślęzak2 Batko and Ślęzak Journal of Big Data (2022) 9:3 , Springer Open.

[3] Big Data Analytics in Medicine and Healthcare Blagoj Ristevski1 / Ming Chen2 Journal of Integrative Bioinformatics. 2018; 20170030.

[4] Big Data Analytics for Healthcare Recommendation Systems Muhib Anwar Lambay , Dr. S. Pakkir Mohideen , B. S. Abdur Rahman , IEEE ICSCAN 2020 , ISBN 978-1-7281-6202-7

[5] Big data analytics: Understanding its capabilities and potential benefitsfor healthcare organizations YichuanWanga,⁎, LeeAnn Kung b, Terry Anthony Byrd aTechnological Forecasting & Social Change126 (2018) 3–13 ScienceDirect

[6] A Mini-Review of Machine Learning in Big Data Analytics: Applications, Challenges, and Prospects Isaac Kofi Nti*, Juanita Ahia Quarcoo, Justice Aning, and Godfred Kusi Fosu. BIG DATA MINING AND ANALYTICS , ISSN 2096-0654 01/06 pp 81 – 97 ,Volume 5 , Number 2, June 2022 , DOI: 10.26599/BDMA.2021.9020028.

[7] Big Data Analytics in Healthcare — A Systematic Literature Review and Roadmap for Practical Implementation Sohail Imran, Tariq Mahmood, Ahsan Morshed, and Timos Sellis, Fellow, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, VOL. 8, NO. 1, JANUARY 2021

[8] A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using N. Ahmed, A. L. C. Barczak, T. Susnjak, and M. A Rashid, HiBench, J , Big Data, vol. 7, no. 1, p. 110, 2020.

[9] Comparative study between Hadoop and Spark based on Hibench benchmarks, in Proc. Y. Samadi, M. Zbakh, and C. Tadonki,2nd Int. Conf. Cloud Computing Technologies and Applications, Marrakech, Morocco, 2016, pp. 267–275.

[10] Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks, Concurrent Computing , Y. Samadi, M. Zbakh, and C. Tadonki, Pract. Exp. vol.30, no. 12, p. e4367, 2018.

[11] https://www.altexsoft.com/blog/hadoop-vs-spark/ date visited 06/11/2022.

[12] Practical Data Science with Hadoop® andSparkDesigning and Building Effective Analytics at Scale OferMendelevitch, Casey Stella, Douglas EadlineCopyright © 2017 Pearson Education, Inc (Book).

[13] https://www.bmc.com/blogs/aws-vs-azure-vs-google-cloud-platforms/ date visited 8/11/2022.

[14] "Who Owns the Data ?", Open Data for Healthcare, P. Kostkova , vol.4. 17 February 2016.

[15] A survey of machine learning for big data processing, EURASIP J . Adv. Signal Process. J. F. Qiu, Q. H. Wu, G. R. Ding, Y. H. Xu, and S. Feng, vol. 2016, no. 1, pp. 67, 2016.

[16] Big data for secure healthcare system: A conceptual design, Complex Intell. Syst, B. K. Sarkar, vol. 3, no. 2, pp. 133–151, 2017.

[17] Big data analytics: A survey, J . Big Data. C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, vol. 2, no. 1, p.21, 2015.