IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 11, 2018, accepted November 7, 2018, date of publication December 21, 2018,
date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2889180

BIG DATA for Healthcare: A Survey

SAFA BAHRI[1], NESRINE ZOGHLAMI[1], MOURAD ABED[2], AND JOÃO MANUEL R. S. TAVARES[3]

[1]LTSIRS Laboratory, University of Tunis El Manar, Tunis 1002, Tunisia
[2]Automatic, Mechanic and Human IT Laboratory, University of Valenciennes and Hainaut-Cambrésis, 59313 Valenciennes, France
[3]Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Unversidade do Porto, 4200-465 Porto, Portugal

Corresponding author: João Manuel R. S. Tavares (tavares@fe.up.pt)

This work was supported in part by the Program of ERASMUS+. The work of J. M. R. S. Tavares was supported by the Fundo Europeu de Desenvolvimento Regional through the Programa Operacional Regional do Norte (NORTE2020) (SciTech—Science and Technology for Competitive and Sustainable Industries) under Project NORTE-01-0145-FEDER-000022.

ABSTRACT Recently, the massification of new technologies, which has been adopted by a large majority of the world population, has accumulated a tremendous amount of data, including clinical data. This clinical data have been gathered up and interpreted by medical organizations in order to gain insights and knowledge useful for clinical decisions, drug recommendations, and better diagnoses, among many other uses. This paper highlights the enormous impacts of big data on medical stakeholders, patients, physicians, pharmaceutical and medical operators, and healthcare insurers, and also reviews the different challenges that must be taken into account to get the best benefits from all this big data and the available applications.

INDEX TERMS Big data, big data analytics, healthcare.

I. INTRODUCTION

Since the massification of the most recent technological trends, like social networks, wearable devices, mobile devices and Internet Of Things, data is being continuously generated in various forms at an unprecedented scale from multiple sources. This massive amount of data, along with the opportunities and possibilities gained from its analysis and the challenges it raises in terms of storage, processing and analysis, has led to the appearance of a new terminology called ''Big Data.''

Big Data has been used by many researchers in diverse fields to support their conclusions and findings. For example, in the transport sector, Big Data Analytics technologies were used in order to improve the service quality, traveler satisfaction and management process, and can suggest ways to optimize customer complaints services [28]. In [29] the effectiveness of Big Data for monitoring smart grid operations is emphasized. The work in [48] is focused on the impact of Big Data Analytics in optimizing airline routes. Also, Big Data has been used in the field of education, where it can play a role in influencing student engagement and behavior [47]. The various applications of Big Data developed between 2010 and 2016 for supply chain management are identified in [6] with the prime objective to convince industry about the effectiveness of Big Data. In addition, the work in [31] identified diverse applications of Big Data in the agriculture sector. By gathering Big Data, organizations can improve customer satisfaction by developing products or services that match their needs and desires. Also, they can enhance their productivity and operational efficiency through resource optimization [27]. Furthermore, the Healthcare sector is considered as one of the main sectors making an evolutionary breakthrough by adopting Big Data techniques and technologies. Indeed, digitalization of medical data has increased tremendously and huge amounts of data are generated at ever increasing rates and in miscellaneous formats, including structured, semi-structured and unstructured datasets. The Institute for Health Technology Transformation [21] announced that the medical Big Data in the U.S reached the zettabyte scale and may soon come to the yottabyte dimension. The main features of Big Data are its diversity and the surprising number of data sources; however, these can be categorized into five groups [7], [21]:

- Large-scale enterprise systems: encompasses data relative to enterprise information systems;
- Online social graphs: is graphic data that illustrates personal relations of social network users;
- Mobile devices: are the main contributors to the Big Data phenomenon; approximately 6 billion smartphones are in use worldwide;
- Internet-Of-Things: By the widespread use of sensors in many organizations, objects are connected with each other and with humans giving rise to huge datasets;

- Open data/public data: which contains data from public and private organizations.

As reported by The McKinsey Global Institute, the USA Healthcare institutions might be able to generate more than US $300 billion in value every year if they applied Big Data creatively and efficiently. Two thirds of these savings would be gained from reducing the healthcare expenditure [49]. Therefore, Big Data in the Healthcare sector is becoming more and more important due to its utility and consistency in, for example, development of Health Recommender Systems, Knowledge Discovery Systems, implementation of Clinical Decision Support Systems as well as Disease Prediction Systems.

This article reviews the state of the art for the application of Big Data in the healthcare sector. Initially, an overview of Big Data and its features are presented. Then, the main aspects of Big Data processes and technologies are discussed. Afterwards, relevant applications of Big Data are identified in the healthcare area. Next, Big Data Analytics are discussed in general terms and especially for the healthcare sector. The article ends with a review of the challenges that were identified in this study, followed by the conclusions (Figure 1).

FIGURE 1. The topics addressed in this article.

II. BIG DATA: DEFINITION AND CHARACTERISTICS

During this last decade, data has been growing exponentially in an unexpected way. As reported by the International Data Corporation (IDC), the volume of data is expected to grow from some zettabytes in 2010 to 163 zettabytes in 2025 (Figure 2). For that, data storage capacity has climbed from megabytes to exabytes and is expected to reach zettabytes per annum in the next few years. Formerly, data was presented in structured formats and stored in relational databases arranged in rows and columns, with limited sources related to internal operations. However, nowadays, many researchers believe that most of this data is unstructured [7], [34], [36], [40] and the use of non-relational (NoSQL) databases is necessary for its management. This data can be categorized into web social media data, machine-to-machine data, transaction data, biometric data and human-generated data [21], [25], [40]:

- ***Social media data*** is acquired from across interactions, tweets and posts in social networks such as Facebook, LinkedIn, YouTube, and Twitter;
- ***Machine-to-Machine data*** is extracted from machine sensors, meters and other devices;

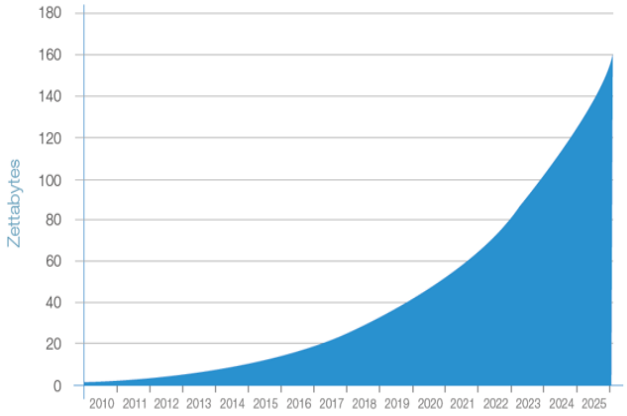

FIGURE 2. The evolution of the data volume between 2010 and 2025 (source: IDC's Data Age 2025 study, April 2017).

- ***Transaction data*** is recovered from fingerprints, genetics, handwriting and medical images;
- ***Human-Generated data*** is selected from prescriptions, emails, messages, documents and Electronic Medical Reports;
- ***Web data*** contains clickstream data generated by internet browsers.

This notable evolution in the growth of data has given arise to this new concept called: "Big Data." To begin with, Big Data is a complex dataset that has prominent influence on the ability of the traditional warehouses to store, handle, manage and analyze it [6]. A formal definition of Big Data was given in [1]: "*Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to acquire specific technology and analytical methods for its transformation into Value*." On the other hand, the McKinsey Global Institute defines Big Data as "*datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze*." For instance, an international discount retail chain (WalMart, USA) stored in 1999 about one terabyte of data. In 2012, it had the capability to store 2.5 petabytes of data derived from customer transactions. Indeed, nowadays, considerable amount of data is generated through the widespread use of mobile devices, sensors, generated-data machines and social media networks. According to [26], the data today is mainly extracted from five sources: Mobile Devices, Internet Of Things, Open and Public Data, Enterprise Information Systems and Social Networks.

A well accepted definition for Big Data nowadays was proposed in [10]: it is a dataset characterized by three Vs (Volume, Variety, Velocity):

Volume: is related to the huge volume of data acquired from various sources that can range from terabytes to exabytes or more [10], [12], [34], [36]. According to IDC, common data volume is going to rise from 1.8 zettabytes to 40 zettabytes between 2011 and 2020 [54]. Therefore, there is a need to manage this data in a parallel way to gain insights.

Variety: it refers to the heterogeneity and diversity of data extracted from several sources like web sites, social media

networks, Electronic Medical Records, Journal Documents, and Video. In fact, Data can be presented in many types, i.e. structured, semi structured and unstructured datasets, having different formats, including image, text and video. Hence, various interpretations and meanings can be extracted from the same dataset [10], [12], [34], [36].

Velocity: describes the rate of data generation that has become time sensitive and frequently needs to be handled and processed in real time. Indeed, many data sources such as sensors generate data that are constantly updated and need to be followed in real time [12], [15], [34], [36].

In addition to the aforementioned features of big data, many researchers have added new features due to the numerous numbers of applications available. Indeed, the initial 3Vs proposal has been extended to 4Vs [3], [26], [42], [50], 5Vs [33], [39], 6Vs [40], 7Vs [24] and 10Vs [18], [36]. An updated list of the commonly adopted Vs is presented in Table 1.

TABLE 1. VsOF BIG DATA.

Vs of Big Data	Meaning
Value	A rigorous use of Big Data Analytics techniques and technologies can extract substantial values, like customer behavior, business performance, rentability and target customer.
Velocity	The speed of data generation.
Veracity or Verification	The trustworthiness and consistency in data as well as quality of data sources are required for accurate analyses and good decision-making.
Variability	From the same data, several interpretations can be found.
Validity	The trustworthiness of data related to a specific application.
Viscosity	This feature is related to velocity. It characterizes the latency in data transfer between the data source and destination.
Volatility	It refers to the validity and storage duration of data.
Visualization	The use of effective visualization tools to present key information and to facilitate the extraction of valuable insights from large amounts of data.
Virility	The measure of how data is diffused to other users and applications.
Valence	The connectedness of data.

III. BIG DATA: PROCESSES AND TECHNOLOGIES

A. BIG DATA CHAIN VALUE

In order to leverage value from this considerable volume of varied data, a four-step process must be followed (Figure 3). Accordingly, a low density set of raw data is processed and analyzed to assist decision maker in their decisions and projects.

1) BIG DATA GENERATION

As discussed in the first section, a tremendous amount of data is being generated from various sources, which includes internal data from company information systems, IoT data, internet data and bio-Medical data. Internal company data encompasses data related to the supply chain, such as production data, quality data, inventory data, sales data and administrative data, including human resources data.

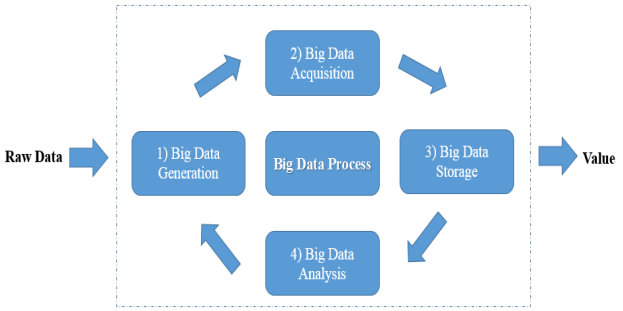

FIGURE 3. Big data chain value: from data collection to data exploitation.

Internet data includes data related to internet search, click stream data, comments and likes, log files and messages. IoT data is related to data generated from devices equipped with sensors and connectivity. Bio-Medical data include data such as genes and drugs data and clinical data [44], [45].

2) BIG DATA ACQUISITION

This second step is usually subdivided into three sub-steps: Data Collection, Data Transmission and Data Pre-Processing:

a: BIG DATA COLLECTION

Big Data Collection is defined as the acquisition and retrieval of unlimited raw data, which can be structured, semi-structured, or unstructured, from several sources using computational techniques and technologies. According to many authors, Big Data sources can be classified into four categories: Information Systems, Mobile devices, Internet Of Things and Open Data [34], [36], [44]. Information System is considered as a centralized data warehouse containing all the information about the activities of an organization. Mobile devices such as smartphones, tablets or PC's generate, a considerable amount of mobile data from the installed applications. Open Data is the large amount of extractable data from, for example, web pages, forums, and journal articles. Internet Of Things encompasses different interconnected devices with embedded sensors able to provide stream and updated data controlled across the internet network.

b: BIG DATA TRANSMISSION

Big Data transmission is related to the transfer of data from data sources into storage management systems for data processing and analysis [44].

c: BIG DATA PRE-PROCESSING

This step ensures efficient and enhanced data for storage and analysis. In fact, collected data must be pre-processed and enhanced by eliminating redundant, noisy, incomplete and useless data leading to a decrease in the storage requirements and an improvement in terms of analytic accuracy. Also, acquired data with low-density needs to be integrated with other data to gain additional value [44].

3) BIG DATA STORAGE

This is the use of databases that can handle a large amount of data with different types and formats for further analysis and processing as well as guaranteeing data security, availability and reliability. Previously, data sources were relatively limited; hence, the Volume, Variety and Velocity of the data were notably smaller, which justified the use of a relational database management system (RDBMS). Currently, with the widespread use of the internet, there is a need to use convenient and efficient data warehouses for processing data. In fact, the data storage equipment is becoming increasingly more important and is considered as the main expense by various institutions [44].

4) BIG DATA ANALYSIS

Big Data Analysis is the most important and critical step in Big Data Chain Value, where value is generated as an output. This is defined as the application of techniques and technologies to mine and extract valuable insights and hidden information from large amounts of processed and stored data [44].

B. BIG DATA TECHNOLOGIES

Formerly, analysts used relational databases and data warehouses to manage and process structured data of limited size. However, according to the commonly adopted Vs of Big Data, traditional tools are inefficient and unable to handle tremendous volumes of data and extract valuable insights from them.

In order to overcome the low performance and the complexity encountered by using traditional technologies, many frameworks and tools based on new distributed architectures along with high memory capacity and processing power have been developed. Accordingly, [37] defined Big Data technologies as: "*a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis*."

Big Data Technologies involve commercial and open-source software and services for storage, analyzing, querying, access, management and processing of data.

Apache Hadoop is a well-known software foundation that offers a collection of open source frameworks designed to support the collection, pre-processing, storage and mining of considerable amounts of data. As presented in Figure 4, this goal is achieved via a master-slave architecture that comprises a unique Name Node and a set of Data Nodes. In this architecture:

- ***NameNode***, which is considered as a master, is responsible for the job scheduling across the cluster. It contains the metadata, which is a data that describe the other data;
- ***A Secondary NameNode*** is a backup *Namenode* that stores all information about the master useful in case the system crashes;
- ***DataNodes*** are slaves that act as executors of all tasks required by the master *Namenode*.

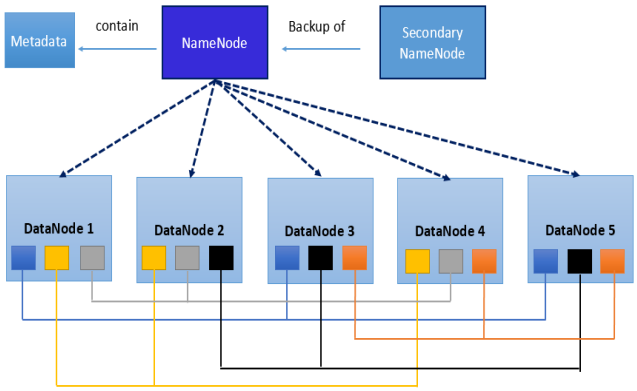

FIGURE 4. Hadoop architecture.

Modern technologies frequently used throughout Big Data processing are:

1) DATA COLLECTION

Sqoop, which is a combination of SQL and Hadoop, is an open source framework that works on top of the Hadoop Distributed File System (HDFS). It is a command-line interface that ensures the import and export of data between HDFS and relational databases, such as Enterprise Data Warehouses, Oracle, Postgres and Teradata (Figure 5). Sqoop presents several benefits for data extraction, such as fast performance and optimized exploitation of resources, and offers excessive storage to other systems and load processing [34], [17].

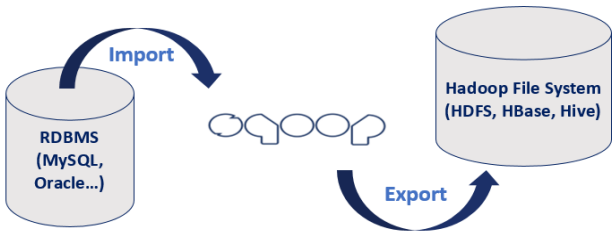

FIGURE 5. Sqoop tool.

Flume is a distributed and reliable open source service designed to assemble, aggregate and transfer huge amounts of batch files, log files and streaming data from external machines to HDFS for storage. It presents a simple and flexible architecture based on continuous streaming data flows. Flume has many advantages such as fault tolerance, robustness and horizontal scalability. An extensible data model is used to process massive distributed data sources [17], [34].

Kafka is an open source framework developed by Apache Software foundation. It is able to collect data from many sources, including data warehouses and social media networks at the same time due to its distributed system and high throughput. LinkedIn and Wikipedia are the main users of Kafka and its benefits. It is written in Scala and characterized by its scalability and fault tolerance [17]. Kafka architecture is composed by Producers, Brockers and Consumers.

Brockers contain Topics, partitions, replicas and offsets. Producers, like Facebook and Twitter, write data to Brockers and Consumers read data from them. Data is stored in Topics that are split into partitions, which are replicated for data security.

Chukwa is a data collection system designed to monitor large distributed systems. It works on the top of the HDFS. It relies on HDFS to collect data from multiple sources and MapReduce to analyze the acquired data. It is known for its scalability and robustness, and offers a friendly user interface to display, monitor and analyze data [17].

2) DATA PROCESSING

MapReduce is a programming model that makes the processing of massive amount of data simpler and faster through its efficient and cost-effective mechanisms. As shown in Figure 6, this framework has three main functions:

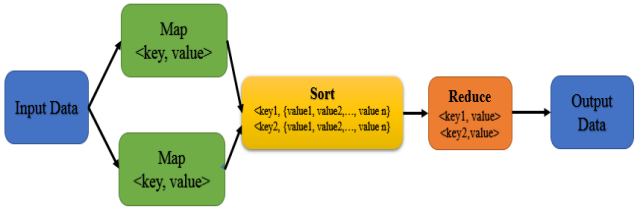

FIGURE 6. MapReduce pipeline.

- ***Function Map***: to ensure that the input data is broken down into independent key/value pairs;
- ***Function Shuffle or Sort***: key/value pairs are collected, stored and then grouped by keys. The output of this function is a collection of keys with associated values;
- ***Function Reduce***: parallel aggregation of pairs according to a predefined program. The outputs are sets of key/value pairs stored in the output file of the MapReduce system.

A well-known application of the MapReduce framework called ''Word Count'' is presented in Figure 7.

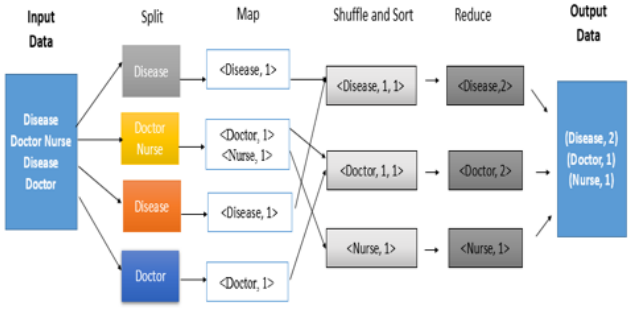

FIGURE 7. A MapReduce application called ''Word Count'' that reads text files in a distributed manner and determines the number of occurrences of each word.

The MapReduce framework also has a master-slave architecture (Figure 8):

- ***JobTracker*** (Master Node): is responsible for the distribution and assignment of tasks for the Slave Nodes;

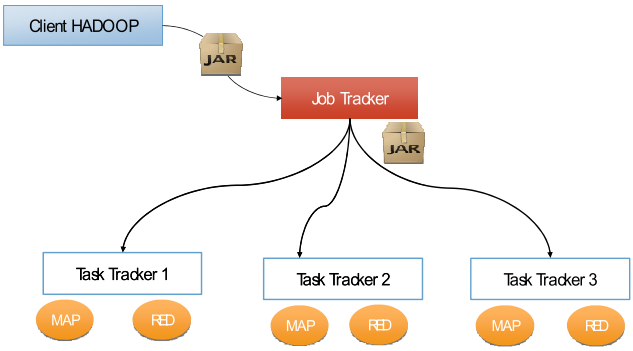

FIGURE 8. MapReduce architecture.

- ***Task Tracker*** (Slaves Nodes): perform tasks required by the JobTracker and supervise their execution.

C++, Python or JAVA are programming languages useful for developing the MapReduce programming model. Many uses of this technology are due to its fault tolerance and scalability. In case of machine failure, a supplementary machine takes care of the node failures [8], [17].

YARN (Yet Another Resource Negotiator) is more generic than MapReduce. It is an advanced resource manager working on top of the HDFS and ensures the parallel execution of various applications. In addition, it handles both batch and stream processing. This framework is also known by its scalability and security. In addition, YARN uses dynamic allocation of system resources, which allows which allows it to increase its exploitation resources. YARN has a master-slave architecture like the MapReduce framework (Figure 9). In fact, the resource manager that operates as a master, manages the assignments of jobs around the cluster. The node manager is a generalized task tracker providing computational resources such as containers, and manages processes running in those containers. A container ensures the execution of the applications-specific process with a constrained set of resources. An application master is in charge of managing the required resources of individual applications. It schedules tasks and assess their progress [8], [17].

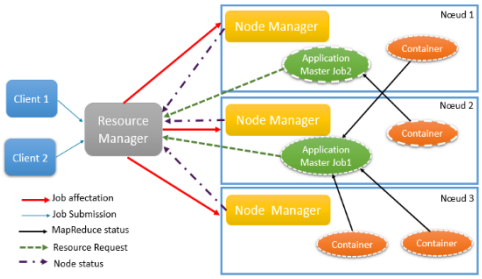

FIGURE 9. YARN architecture.

Storm: is an open source framework designed for distributed real time computations. Unlike HDFS that is targeted for batch processing, this tool is able to handle stream data. It is characterized by its high scalability, fault tolerance

and efficiency. Storm is used, for example, in real time analytics, ETL (Extract, Load, Transform) operations, online machine learning, and continuous computations [8], [17].

Flink: is an open source framework able to process stream data due to its distributed architecture.

Spark: is an in-memory cluster computing technology developed for fast computing of data through its sophisticated libraries:

- *Spark Streaming*: it allows the collection and processing of data in real-time;
- *Spark SQL*: it is able to execute SQL queries;
- *Spark MLlib*: is a Machine Learning library useful and powerful in solving machine learning problems. It contains different machine learning algorithms related to classification, regression, clustering and optimization;
- *Spark R*: it contains the same functionalities of the programming language for statistical computing and graphics "R." However, processing time is reduced thanks to its distributed architecture;
- *Spark GraphX*: it is devoted to the processing of graphs.

Giraphe: represents an interactive graph processing with high scalability. It is mostly used by Facebook to interpret social graphs and connections between subscribers [8].

Zookeeper: is a distributed service designed for the synchronization of configurations across a cluster. It ensures the high availability of data [17], [27].

Oozie: is an open-source job coordinator that executes and manages job flows in the Hadoop system. Oozie is a scalable, reliable and extensible system [27].

3) BIG DATA STORAGE

To ensure the storage of a huge volume of data, Hadoop uses the distributed data storage system HDFS and a non-relational database named HBase:

HDFS is one of the primary components of a Hadoop cluster, i.e. a set of connected computers, that can support up to hundreds of nodes in a cluster. It is cost effective and has a reliable storage capability, high scalability as well as fault tolerance. In addition, HDFS can handle both structured and unstructured data. HDFS is designed for batch processing of high latency operations. In fact, it stores data in 64 or 128 byte capacity blocks. In order to avoid data losses, blocks of the same file are replicated three times and stored across the cluster in three different servers. HDFS has a master-slave architecture (Figure 4) that is commonly adopted due to its capacity to reduce network congestion and increase system performance by performing the computations near the data storage locations [17], [12].

HBase is an open source project built on top of HDFS designed for low latency operations. This non-relational database, developed posterior to Google's Big Table, has the potential to host very large tables with billions of rows and millions of columns. Unlike a row oriented relational database that stores together all columns of a row, HBase is a columnar database management system that stores data in columns to ensure easier access to data. As for HDFS, this database also has a master-slave architecture. The master node manages the cluster, and the slaves perform the required operations on the available data. HBase is a flexible, distributed and scalable database, and has the capability for real-time queries, automatic and configurable partitioning of data to facilitate data processing and analysis [17].

HCatalog: is a table and storage management system for Hadoop that enables users with different data processing tools to read and write data on the grid more easily [17], [27].

Avro: is an open source framework designed by Apache Hadoop that offers two services for developers: Data serialization and Data exchange [17], [27].

4) BIG DATA ANALYSIS

Pig: is an open source platform designed by Yahoo to analyze large datasets that are considered to be data flows. In order to develop data analysis programs, Pig uses a high programming language called "Pig Latin." Accordingly, to analyze data, developers must write Pig Latin scripts then convert them into MapReduce tasks using the component Pig Engine. Apache Pig presents many advantages. First, due to its multi-query approach, the length of codes is reduced. Second, Pig Latin substitutes the use of Java, which is traditionally seen as more complicated, when coding MapReduce jobs. Indeed, Pig Latin is similar to SQL language and is easy to learn. The only difference is that Pig Latin is able to process semi-structured and unstructured data. Finally, Pig is characterized by its interactive environment and can process massive amounts of data due to its distributed architecture [17], [45].

Hive is a data warehouse system created by Facebook in order to facilitate the use of Hadoop. The collected data is stored in a structured database, comprehensible to all users. Apache Hive database is managed through a HQL language having the same syntax as SQL language. HQL transforms queries into MapReduce jobs processed as batch tasks. Like Pig, Hive has an interactive interface with a diversity of functions useful for data analysis. Unfortunately, Hive is mostly used for structured data [17], [45].

Mahout is an open source library designed for machine learning and data mining. It works on the top of HDFS in order to execute algorithms via MapReduce. It helps developers access their own libraries for clustering, collaborative filtering, categorization and text mining. It is scalable and can be executed in a distributed mode [17], [27].

To conclude, the Hadoop Ecosystem (Figure 10) contains very powerful tools able to collect, process, store and analyze a large amount of varied data coming from several sources generated at a high rate. This vast potential is due to its scalability, fault tolerance, flexible scheduling and resource management, high level and simplified programming model, distributed architecture, real-time processing, in-memory processing and high throughput.

The next section presents pertinent applications of Big Data, especially in the Healthcare sector.

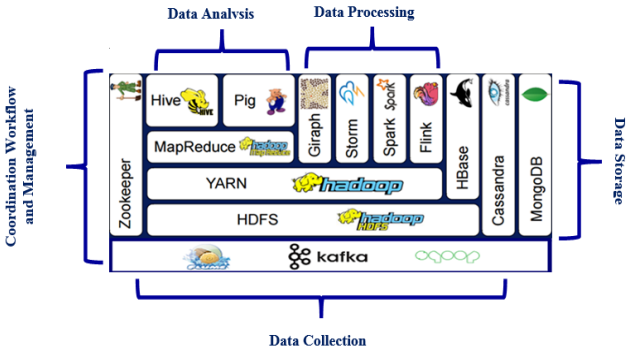

FIGURE 10. Hadoop ecosystem.

IV. BIG DATA APPLICATIONS IN HEALTHCARE SECTOR

Healthcare is among the sectors generating a massive amount of data characterized by its high velocity and variety such as laboratory data, medical prescriptions, appointments, machine generated data, insurance data, and administrative data. In fact, in USA, only clinical data reached 150 exabytes in 2011. According to predictions, the volume of clinical data will soon reach zettabytes or even yottabytes [21], [40]. The authors in [7] estimate that 90% of all clinical data are unstructured, such as prescription notes, lab results, and ECG data. According to the authors in [19], Big Data capability in the healthcare sector is defined as: "*The ability to acquire, store, process and analyze large amounts of health data in various forms, and deliver meaningful information to users that allow them to discover business values and insights in a timely fashion.*" Hence, there is a need to analyze this Big Data in order to:

- improve patient satisfaction and quality of care;
- reduce expenditures, which reached 17.9% of the gross domestic product in 2010 [21].

The authors in [43] suggest that the collection of clinical data should adopt the best strategies and recommendations in order to overcome the usual performance gaps. This data is generated from many sources that encompasses Electronic Health Records, medical imaging data, genetic data and prescription notes [34]. Some examples of pertinent Big Data Applications in the Health Sector are the following.

A. HEALTHCARE MONITORING

A deep analysis of the healthcare data can help care providers manage symptoms of patients online and adjust prescriptions [17]. For instance, with the development of wearable sensor devices, like Apple Watch and Sports bracelets, information related to physical health checkups, including blood pressure, height, weight, blood-glucose levels and blood-calcium levels, can be constantly monitored in order to give a detailed vision about the condition of the patient's health. These indicators help physicians monitor patients and consequently unnecessary visits to the doctor can be avoided and at the same time, patients have the impression that they are more independent and yet become more aware concerning their healthcare status [15], [36]. Furthermore, smart dispensers are used to detect if drugs are being taken regularly at the right time. In the case of non-medication, the practitioner can intervene to get patients properly medicated. Besides, Big Data analytics help physicians to avoid medication errors due to drug interactions or incorrect dosages, which could easily lead to a critical situation. The Big Data allows the physicians to assess a patient's records in full including the recommended medications. In USA, about one million injuries occur due to prescription errors leading to thousands of unnecessary deaths. Therefore, an analytical solution called "MedAware" has been developed. A real-time analysis of medical prescriptions can detect a wide range of errors with high rates of precision (90%). Anticipating clinical errors can reduce unnecessary hospitalizations and re-admissions as well as excessive lengths of hospitalization stays [23].

Furthermore, Common Sensing, a company that use Big Data technologies to ensure a follow-up of treatment given to diabetic patients, developed a replacement cap named "GoCap" for prefilled insulin pens able to register the dosage of insulin taken daily and the exact time when it was administrated. Then, Bluetooth technology is used to transmit the data collected to a mobile phone or a connected glucometer. This stream data might be transferred to the care providers who would be able to detect potential healthcare issues and to interrupt treatment in case of an emergency [16].

B. HEALTHCARE PREDICTION

Social networks, such as Facebook and Twitter, can help establish healthcare social networks. For example, patients who suffer from chronic diseases can share their own experience with other patients or doctors. This sharing helps them to benefit from a broad range of experiences and expertise [30], [40]. An integrative healthcare system called "GEMINI" was developed to process and analyze a large amount of variable and complex healthcare information. First, for each patient, data is collected from structured sources, including patient demographics, lab test results and medication history, and from unstructured sources, for example, prescriptions. This data is then stored in a patient profile graph that presents a broad view of the patient's state of health. Then, analytics algorithms such as classification and clustering algorithms are used to extract valuable insights useful for administrative and clinical purposes as well as predictive analytics [41].

In addition, big genomics data analytics intervene in healthcare predictions by measuring the changes in DNA mutations and detecting the molecular responsible for the appearance of diseases [34].

Moreover, in order to reduce the healthcare costs due to high rates of patients that suffer from chronic diseases, such as diabetes, doctors use big data predictive analytics to foresee high-risk patients and to offer to them customized care [42]. To ensure a therapeutic follow-up for asthmatic patients, an American company named Asthmapolis has

developed sensors that are positioned in the top of an inhaler in order to track and follow the inhaler usage. Data related to the place and time of inhaler's use is collected in real time via global positioning system (GPS). In the case of asthma attack, recorded data is transferred to a web site accessible by the patient and doctor via a smartphone or computer. On one hand, the main goal of the sensors is to help patients make a good decision about visiting different places or not as well as predicting asthma attacks. On the other hand, the data aggregation guides the practitioner to develop a customized care plan for the patient, identifying epochs with a high probability of an asthmatic crisis and anticipating them by either increasing or decreasing the drug dosage [16].

The authors in [44] selected 102 patients from 1000 that suffer from metabolic syndrome, to follow-up on their recovery. Analysts collected data from 600 laboratory tests and 180 claims. Moreover, a customized treatment plan was elaborated from patients' health records. This application produced encouraging results: The morbidity rate might be reduced by 50% over the 10 next years. Many alternative solutions were deduced: Prescription of statins, loss of weight and reduction in the total triglyceride rate in the body if the blood sugar level exceed 20%.

In order to follow the propagation of epidemics around the world, Google has developed two real-time surveillance applications: Google flu trends and Google dengue trends. Data were collected from internet searches where, in fact, people were going to Google search engine to find information about symptoms, drugs, side effects, etc. The results confirmed that Google Flu Trends over forecasted the prevalence of flu by 140% [12].

C. PERFORMANCE ENHANCEMENT

In order to maximize the performance of Emergency Rooms (ER) and decrease crowding in ER, King Faisal Specialist Hospital and Research Center was successfully managed by a project based on clinical big data analytics. Data relating to the emergency department was extracted from the data warehouse of the hospital. The analysis of the data introduced changes in the workflow of the ER which led to positive results. In fact, the ER waiting time reduced from 140 min in 2014 to 62 min in 2016. The treatment time decreased from 17.5 h in 2014 to 10.8 h in 2016. The ER Length Of Stay varied from 20 h in 2014 to 12 h in 2016. This showed the effectiveness of big data analysis in identifying areas of insufficiency, and in recommending valuable solutions to positively improve performance [43].

D. RECOMMENDATION SYSTEMS

In general, recommendation systems are software tools and techniques developed to propose a set of suggestions for a product or a service, which help users make better decisions. Currently, in the healthcare sector, recommendation systems are increasingly used in order to provide medical recommendations for drugs, diagnoses, and treatment plans.

The authors of [46] developed a clinical recommendation system useful for patients to obtain reliable recommendations of care providers in reference to their own health status. In addition, patients are able to contribute to the enhancement of the performance of the system by adding their private notes and evaluations about physicians based on data for different health conditions. However, due to the sensitivity of such data, it must be protected from dishonest and malicious users. Moreover, according to care providers, this system ensures the preservation of their reputation. As an output, this system presents, for each patient, a list of best-ranking physicians.

Furthermore, a collaborative-filtering method is considered a type of recommendation system that forecasts viewpoints of users about an article referring to the preference of a large group of users [13]. This technique was then applied in the healthcare sector, when a Collaborative Assessment and Recommendation Engine (CARE) for disease risk prediction was created. The first visit of a patient to a doctor provides clinical input data related to the medical history of the patient. And for each disease j, analysts have to find patients who suffer from this disease based on health record data. The application of collaborative filtering generates $p(a, j)$ that denoted the probability that a patient a will develop disease j in the future. Finally, for each patient, the system provides the physician with a sorted list of potential diseases ranked in order from highest to lowest risk. This framework is effective for decreasing readmission rates, making consistent predictions and enhancing quality of care ratings. To ensure its feasibility and efficiency, the system was validated on a Medicare database of 13 million patients who made 32 million medical visits over 4 years.

E. HEALTHCARE KNOWLEDGE SYSTEM

A knowledge System is defined as the combination of information, data and physician expertise in order to present alternatives to potential emergency situations and to support clinical decision-making and diagnosis. The authors in [34] suggested a healthcare knowledge system based on four Big Data sources: Electronic Health Records (EHR), Clinical Notes, Genetic Data and Medical Imaging Data. EHR data includes structured data, for example, laboratory data and billing data, and unstructured data such as medication records. Laboratory data is useful for diagnosis and health monitoring. Billing data includes various codes giving access to laboratory results, clinical records and symptoms. Medication records encompass a variety of data useful for disease diagnosis and drug recommendations. Clinical Notes are unstructured data aiming to identify common or well-known illnesses. Genetic Data are a huge volume of data that is used in the analysis of changes in gene sequences. And unstructured medical imaging data contains different image data useful for treatment, diagnosis and prognosis.

F. HEALTHCARE MANAGEMENT SYSTEM

The authors in [9] proposed a smart Healthcare Management System named "*DataCare*", to be used at hospitals or

healthcare centers, with the aim to increase patient satisfaction by monitoring clinical Key Performance Indicators (KPIs) and identifying unexpected situations. The architecture of the system has three main modules: Data Retrieving and Aggregation, Data Processing and Analysis, and Data Visualization. In the first module, data is collected and aggregated via AdvantCare software which is used for supervising the communication between patients and physicians. In the second module, the authors decided to use Apache Spark to process the huge amounts of streaming data due to its high scalability and fault-tolerance. Data was then stored in a MongoDB database for further analysis. The stored data was analyzed in order to extract valuable insights for healthcare predictions, clinical recommendations and alerts. In fact, DataCare is capable of forecasting KPIs in the future based on actual data. Moreover, the framework is able to generate early and real-time alerts if an indicator exceeded its authorized value delimited by thresholds. In addition, DataCare provided recommendations aiming to enhance the quality of care. DataCare has 52 rules designed by physicians based on their expertise and knowledge. Finally, the Visualization module displayed all the information on dashboards to ensure more accurate and efficient interpretations. DataCare was validated at a medical centre in Spain. Expected outcomes were obtained with interesting conclusions.

From these diverse applications in the healthcare sector, one can deduce the potential that Big Data analysis can have on optimizing hospital operations, improving care, decreasing expenditure and readmission rates, saving lives, and improving quality of care [11], [17], [22].

V. BIG DATA ANALYTICS

Big Data Analytics is defined as mining of pertinent knowledge and valuable insights from large amounts of stored data [4]. The key objective of such analytics is to facilitate decision making for researchers, such as offering dashboards, graphics or operational reporting to monitor thresholds and KPIs. This involves using mathematical and statistical methods to understand data, simulate scenarios, validate hypotheses and make predictive forecasts for future incidents. Data Mining is a key concept in Big Data Analytics that consists in applying data science techniques to analyze and explore large datasets to find meaningful and useful patterns in those data. It involves complex statistical models and sophisticated algorithms, such as machine learning algorithms, mainly to perform four categories of analytics: Descriptive analytics, Predictive analytics, Prescriptive analytics and Discovery (Exploratory) analytics. Descriptive analytics turns collected data into meaningful information for interpreting, reporting, monitoring and visualization purposes via statistical graphical tools such as pie charts, graphs, bar charts, and dashboards. Predictive analytics is commonly defined as data extrapolation based on available data for ensuring better decision making. Prescriptive analytics is associated with Descriptive and Predictive analytics. Likewise, based on the present situation, it offers options on how to benefit from future opportunities or mitigate a future risk and details the implication of each decision option. Finally, Discovery (Exploratory) analytics illustrates unexpected relationships between parameters in Big Data [4]. The authors in [24] argue that currently, the output of predictive analytics can benefit from the potential of descriptive analytics through the use of dashboards and scorecard computations.

Big Healthcare Data Analytics (BHDA) is defined as the use of statistical, cognitive, predictive, contextual, and quantitative models for efficient and fast decision making useful for planning, forecasting, resource management, etc. Big Data Analytics helps healthcare stakeholders, medical practitioners, hospital operators, pharmaceutical and clinical researchers, and healthcare insurers, to improve their findings by harnessing their internal and external Big Data [5], [11], [12], [40]. According to medical practitioners, the analysis of patient data, including patient medical history, physicians' notes, laboratory results, and clinical trials data, assists them to track the progress of a proposed treatment plan and to interrupt the plan to make changes if necessary, and consequently unnecessary visits can be eliminated and readmission rates decreased. For hospital operators, Big Data Analytics helps them to allocate resources. For instance, the analysis of location awareness data contributes to optimize the use of expensive healthcare equipment and devices. In addition, pharmaceutical organizations take profit from analytic advantages in the elaboration of marketing strategies. In fact, by gathering and analyzing data such as sales history data, and drug recommendation for each patient and disease, they are able to assess their current market position, which is useful for the definition of strategic priorities. Furthermore, the analysis of patient demographic data, such as age and gender, and clinical data, such as disease and drugs history, the insurer is able to elaborate an appropriate health plan for each patient [30]. In conclusion, Big Data Analytics plays an important role in the enhancement of medical services and increases patient satisfaction. Consequently, it has the potential to improve care, save lives and lower costs.

VI. BIG DATA: CHALLENGES & PERSPECTIVES

A. BIG DATA CHALLENGES

Big Data presents various opportunities in the medical, biomedical and healthcare sectors due to its ability to obtain valuable knowledge useful for improving healthcare organizations, reducing healthcare costs as well as reducing unnecessary visits and readmissions [11]. However, by handling datasets characterized by huge volumes generated at very high speeds and with large diversity, e.g. structured, semi structured, and unstructured data, many barriers and hurdles have to be overcome along the path to create value from data collection to data analysis. Diverse challenges, that can be categorized into five groups, must be overcome to be able to benefit from the advantages of Big Data:

1) DATA COLLECTION

Data reliability is among the criteria for data selection during the collection phase. It is crucial to select the data sources well, considering that they may contain noise, errors, as well as inconsistent or incomplete data. However, due to the enormous diversity of sources, it is becoming a challenge to treat it all and select the best. In addition, during data acquisition, it is necessary to integrate the external data of the organization to the internal data in order to obtain knowledge and updated information about the external environment and to make accurate prediction models. This aggregation is becoming more and more challenging [17].

2) DATA PROCESSING

Data processing aims to generate cleaned, consistent and secured data for efficient and accurate analysis. The first challenge is how to collect, process and store variable data from various types of devices with limited capacity and CPU. The second challenge is how to ensure accuracy and consitency in decision making when aggregating dissimilar data with multiple formats [17]. However, many Big Data processing tools perform poorly with computational uncertainties, unconsistencies and complexities. So, it is becoming a challenge to use convenient techniques and technologies that aim to minimize computational cost processing and complexities. Currently, many well-known organizations require real time data processing in which large amounts of data are promptly executed in real or near-real time to allow fast decision making. Therefore, there is a need to adopt Big Data technologies with high scalability [32].

3) DATA STORAGE

The volume of data collected is increasing dramatically, especially due to the spread and use of new technological trends, such as social media, and remote sensing. In the recent past, developers and analysts were using hard drive disks to store data. Unfortunately, these devices are now not suitable for data storage. Therefore, the first challenge is the usage of apropriate storage mediums with higher input/output speeds. The main goal is to ensure data availability and accessibility for further analysis [32].

4) DATA ANALYSIS

For the extraction of relevant information from a pool of stored data generated from different sources, it is crucial to choose the analytical software and hardware well in order to produce more accurate and valuable outcomes. This requirement is becoming more and more challenging due to diversity of available technologies designed for data analysis. Besides, the variety of data can cause unprecedented challenges for analysts. Indeed, the existing tools are unable to respond in the required time when treating high dimensional data. The next challenge is related to how effectively multivariate data can be analyzed in order to obtain valuable knowledge as an output. Moreover, for an easy and pertinent interpretation and analysis, many researchers deal with the use of graphic vizualization tools capable of summarizing large amounts of data into significative and intuitive graphic or picture formats. Thus, researchers need to use tools with high scalability. However, most of the tools available present many functional limits in terms of scalability and timeliness responses. So, once again it is a challenge to design software or hardware that will lead to to parallel computing and visualization processes to ensure accurate analyses [2], [32].

5) DATA SECURITY

Clinical data are very sensitive data that must be made secure in order to protect data from hackers that are able to use data mining techniques to extract personnal data and make it public. Therefore, it is essential to implement security procedures such as authentification, authorization and encryption to enhance data security. The challenge here is to develop a multi-level security, privacy preserved data model for Big Data [20], [32], [35].

B. BIG DATA CHALLENGES FOR HEALTHCARE

Currently, the healthcare sector is among the sectors that generate tremendous amount of data from multipe sources that can be quantitative, including laboratory test, genes arrays and sensor data, or qualitative, such as demographics and free texts [35]. According to [3], medical data is different from other data. In fact, it is very sensitive and hard to access. Unfortunately, data can be affected by the use of unmanaged data sources such as social networks. Also, any errors in measures or in codes can severly affect the reliability of an analysis. Therefore, data trustworthness, data quality and data consistency are yet another challenge for Big Data. Moreover, clinical data is continously being generated, hence, the use of real-time data streaming tools and technologies are gaining relevance [22].

As previously stated, the majority of the population in the world use social networks to collect updated information as well as to obtain knowledge, to communicate and to carry out personal research. Therefore, a tremendous amount of data will be generated at high speed and in various formats. Thus, it is very interesting and challenging to exploit and integrate this data in order to improve healthcare outcomes. A possible challenge is to elaborate a medical decision system for forecasting the spreading of epidemics in different geographical locations by analysing social media data such as Facebook and Twitter. In order to carry this out, a robust predictive system that considers a set of attributes related to a type of epidemic, e.g. influenza, dengue fever or cholera, must be developed. The past values extracted from social media including comments, likes, and posts, will allow a predictive system to extrapolate these values into the future. The capabilities of Big Data technologies capabilities and the richness, massiveness and variety of social network data can be combined to provide relevant healthcare predictions.

VII. CONCLUSION

This study aimed to emphasize the enormous implications of Big Data Techniques and Technologies on the performance and outcomes of Healthcare organizations. Section 1 presented this novel concept "Big Data" and its evolution over time as well as its Vs: Volume, Variety, Velocity, Veracity, Variability, Validity, Viscosity, Volatility, Visualization, Virility, and Valence. Then, Section 2, described the process of building value from big data. For each step, a list of available Big Data technologies was proposed and detailed. This section showed the potential of technologies in handling and analyzing huge amounts of data extracted from multiple sources and in different formats. Then, in Section 3, big data applications for healthcare found in the literature were classified into five groups: Healthcare monitoring, Healthcare Prediction, Recommendation systems, Healthcare Knowledge system and Healthcare Management System. Based on the reviewed cases, one can confirm the countless opportunities offered by Big Data and its analysis.

The analytical capabilities of Big data techniques and technologies as well as the consistent knowledge and valuable insights that can be derived from stored Big Data are useful for making predictions, recommendations, medical diagnosis, resource allocations and personalized treatment plans. This ability may have a positive effect on the quality of healthcare and its outcomes. Here, Big Data Analytics was classified into four types: Descriptive Analytics, Prescriptive Analytics, Predictive Analytics and Discovery Analytics. Finally, based on the Big Data features identified, along with Big Data Chain Value, many of the challenges that must be tackled, were identified.

REFERENCES

[1] A. De Mauro, M. J. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Rev.*, vol. 65, no. 3, pp. 122–135, 2016.

[2] 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.

[3] C. H. Lee and H.-J. Yoon, "Medical big data: Promise and challenges," *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, 2017.

[4] D. Rajeshwari, "State of the art of big data analytics: A survey," *Int. J. Comput. Appl.*, vol. 120, no. 22, pp. 39–46, 2015.

[5] J. Cortada, D. Gordon, and B. Lenihan, "The value of analytics in healthcare: From insights to outcones," IBM Global Services, Armonk, NY, USA, Executive Rep. GBE03476-USEN-00, 2012. [Online]. Available: http://www-05.ibm.com/ch/gesundheitswesen/pdf/The_value_of_analytics_in_healthcare.pdf

[6] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," *Comput. Ind. Eng.*, vol. 115, pp. 319–330, Jan. 2018.

[7] N. Ilyasova, A. Kupriyanov, R. Paringer, and D. Kirsh, "Particular use of BIG DATA in medical diagnostic tasks," *J. Sci. Commun.*, vol. 28, no. 1, pp. 114–121, 2018.

[8] S. Mazumder, "Big Data Tools and Platforms," in *Big Data Concepts, Theories, and Applications*, Y. Shui and G. Song, Eds. Cham, Switzerland: Springer, 2016, pp. 29–128.

[9] A. Baldominos, F. De Rada, and Y. Saez, "DataCare: Big data analytics solution for intelligent healthcare management," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 4, no. 7, pp. 13–20, 2017.

[10] D. Laney, "3D data management: Controlling data volume, velocity and variety," META Group, Stamford, CT, USA, Res. Note 6, 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[11] A. Kankanhalli, J. Hahn, S. Tan, and G. Gao, "Big data and analytics in healthcare: Introduction to the special section," *Inf. Syst. Frontiers*, vol. 18, no. 2, pp. 233–235, 2016.

[12] V. Rajaraman, "Big data analytics," *Resonance*, vol. 21, no. 8, pp. 695–716, 2016.

[13] V. Shobana and N. Kumar, "A personalized recommendation engine for prediction of disorders using big data analytics," in *Proc. IEEE ICIGEHT*, Coimbatore, India, Mar. 2017, pp. 1–4.

[14] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: A patient-centered framework," *J. Global Inf. Manage.*, vol. 28, no. 3, pp. 660–665, 2013.

[15] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, "A survey on big data market: Pricing, trading and protection," *IEEE Access*, vol. 6, pp. 15132–15154, 2018.

[16] R. Nambiar, A. Sethi, R. Bhardwaj, and R. Vargheese, "A look at challenges and opportunities of big data analytics in healthcare," in *Proc. IEEE Int. Conf. Big Data*, Silicon Valley, CA, USA, Oct. 2013, pp. 17–22.

[17] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big data technologies: A survey," *J. King Saud Univ., Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, 2017.

[18] K. Tiampo, S. McGinnis, Y. Kropivnitskaya, J. Qin, and M. A. Bauer, "Big data challenges and hazards modeling," in *Risk Modeling for Hazards and Disasters*, M. Gero, Ed. Amsterdam, The Netherlands: Elsevier, 2018, pp. 193–210.

[19] Y. Wang, L. Kung, and T. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technol. Forecasting Social Change*, vol. 126, pp. 3–13, Jan. 2016.

[20] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, "Big data security and privacy in healthcare: A review," in *Proc. 8th Int. Conf. EUSPN*, Lund, Sweden, 2017, pp. 73–80.

[21] *Transforming Healthcare Through Big Data: Strategies for Leveraging Big Data in the Healthcare Industry*, Inst. Health Technol., New York, NY, USA, 2013.

[22] W. Raghupathi and V. Raghupathi, "Big Data analytics in healthcare: Promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 3, pp. 1–10, 2014.

[23] *MedAware Raises $8 Million for Software to Reduce Prescription Errors*, Wall Street J., Dow Jones Company, New York, NY, USA, 2017.

[24] U. Sivarajah, M. M. Kamal, I. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017.

[25] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Bus. Horizons*, vol. 60, no. 3, pp. 293–303, 2017.

[26] B. Baesens, R. Bapna, J. R. Marsden, J. Vanthienen, and J. L. Zhao, "Transformational issues of big data and analytics in networked business," *MIS Quart.*, vol. 40, no. 4, pp. 807–818, 2016.

[27] N. Khan *et al.*, "Big data: Survey, technologies, opportunities, and challenges," *Sci. World J.*, vol. 2014, Art. no. 712826, doi: 10.1155/2014/712826.

[28] W.-K. Liu and C.-C. Yen, "Optimizing bus passenger complaint service through big data analysis: Systematized analysis for improved public sector management," *MDPI J.*, vol. 8, no. 12, p. 1319, 2016.

[29] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big data management in smart grid: Concepts, requirements and implementation," *J. Big Data*, vol. 4, no. 13, pp. 01–19, 2017.

[30] V. Palanisamy and R. Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks—A review," *J. King Saud Univ.-Comput. Inf. Sci.*, to be published, doi: 10.1016/j.jksuci.2017.12.007.

[31] V. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Bold, "A review on the practice of big data analysis in agriculture," *Comput. Electron. Agricult.*, vol. 143, pp. 23–37, Dec. 2017.

[32] D. P. Acharjya and A. P. Kauser, "A survey on big data analytics: Challenges, open research issues and tools," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 2, pp. 511–518, 2016.

[33] R. Y. Zhong, S. T. Newman, G. Q. Huang, and S. Lan, "Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives," *Comput. Ind. Eng.*, vol. 101, pp. 572–591, Nov. 2016.

[34] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K. M. Abbas, and R. Sundarsekar, "Big data knowledge system in healthcare," in *Internet of Things and Big Data Technologies for Next Generation Healthcare*, C. Bhatt, N. Dey, and A. S. Ashour, Eds. Springer, 2017, pp. 133–157.

[35] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big data for health," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[36] G. Manogaran, D. Lopez, C. Thota, K. M. Abbas, S. Pyne, and R. Sundarsekar, "Big data analytics in healthcare Internet of Things," in *Innovative Healthcare Sytems for the 21st Century*, H. Oudrat-Ullah and P. Tsasis, Eds. Springer, 2017, pp. 263–284.
[37] IDC Iview. vol. 1142, pp. 1–12, 2011. [Online]. Available: https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
[38] IDC Country Brief. Feb. 2013. [Online]. Available: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf
[39] R. Addo-Tenkorang and P. Y. Helo, "Big data applications in operations/supply-chain management: A literature review," *Comput. Ind. Eng.*, vol. 101, pp. 528–543, Nov. 2016.
[40] S. Shafqat, S. Kishwer, R. ur Rasool, J. Qadir, T. Amjad, and H. F. Ahmad, "Big data analytics enhanced healthcare systems: A review," *J. Supercomput.*, pp. 1–46, 2018, doi: 10.1007/s11227-017-2222-4.
[41] Z. J. Ling *et al.*, "GEMINI: An integrative healthcare analytics system," in *Proc. 40th Int. Conf. Very Large Data Bases*, Hangzhou, China, 2014, pp. 1771–1776.
[42] N. M. S. Kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive methodology for diabetic data analysis in big data," in *Proc. 2nd Int. Symp. Big Data Cloud Comput. (ISBCC)*, vol. 50, 2015, pp. 203–208.
[43] M. Khalifa and I. Zabani, "Utilizing Health Analytics in improving the performance of healthcare services: A case study on a tertiary care hospital," *J. Infection Public Health*, vol. 9, no. 6, pp. 757–765, 2016.
[44] M. Chen, S. Mao, Y. Zhang, and V. C. Leung, *Big Data: Related Technologies, Challenges and Future Prospects*. Cham, Switzerland: Springer, 2014.
[45] A. Bhadani and D. Jothimani, "Big data: Challenges, opportunities and realities," in *Effective Big Data Management and Opportunities for Implementation*, M. K. Singh and D. G. Kumar, Eds. Hershey, PA, USA: IGI Global, 2016, pp. 1–24.
[46] T. R. Hoens, M. Blanton, A. Steele, and N. V. Chawla, "Reliable medical recommendation systems with patient privacy," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, 2013, Art. no. 67.
[47] R. J. Watson and J. L. Christensen, "Big data and student engagement among vulnerable youth: A review," *Current Opinion Behav. Sci.*, vol. 18, pp. 23–27, Dec. 2017.
[48] E. Kasturi, S. P. Devi, S. V. Kiran, and S. Manivannan, "Airline route profitability analysis and optimization using BIG DATA analyticson aviation data sets under heuristic techniques," *Procedia Comput. Sci.*, vol. 87, pp. 86–92, 2016.
[49] J. Manyika *et al.*, *Big Data: The Next Frontier for Innovation, Competition and Productivity*. New York, NY, USA: McKinsey Global Institute, 2011.
[50] A. Intezari and S. Gressel, "Information and reformation in KM systems: Big data and strategic decision-making," *J. Knowl. Manage.*, vol. 21, no. 1, pp. 71–91, 2017.

SAFA BAHRI was born in Tunis, Tunisia, in 1992. She received the National Diploma of Engineering degree in industrial engineering from the National School of Engineering of Carthage, in 2016. She is currently pursuing the joint Ph.D. degree with the LTISIRS Laboratory, National School of Engineering of Tunisia, and the LAMIH Laboratory, University of Valenciennes and Hainaut-Cambrésis. She is developing a healthcare system that predicts the spreading of epidemics through social media data analysis. Her research interests include big data applications in the healthcare sector.

NESRINE ZOGHLAMI received the Diploma degree in electrical engineering and the M.Sc. degree in electronic and telecommunications from the University of Valenciennes, France, and the Ph.D. degree in industrial computer science and automatic from the Ecole Centrale de Lille, France, in 2008. She is currently an Associate Professor in industrial computer science. She took an active part in the national working group ORT of GDR MACS. She was responsible for organizing the program committees of international conferences and workshops, including MHOSI 2005, LT 2006, LT 2007, LT 2009, MSLT 2011, Sysco 2012, ICALT 2013, GOL 2014, ICALT 2014, ICALT 2015, ICIT 2015, ICIT 2016, GOL 2016, IPAC2016, BDWA 2016, ASET 2017, and ASET 2018. She was a Rapporteur of the H2020 proposals for the Research Executive Agency, European Commission, from 2015 to 2016. She has authored or co-authored about 50 publications, communications, and book chapters. She has authored a book *Optimisation Ãă base d'agents communicants des flux logistiques pour la gestion de crise* on supply chain management. Her main research interests include optimization, artificial intelligence, and supply chain management.

MOURAD ABED was the Vice-President (digital) of the University of Valenciennes and the Vice-Director of the Institute of Science and Technology, from 2000 to 2010. He is currently a Professor (Exceptional class) in computer engineering with the University of Valenciennes and a member of the Human Computer Interaction and Automated Reasoning Research Group, Automatic, Mechanic and Human IT Laboratory. He is also the Director of the Program of Master of Science and Technology Studies, a European Project Coordinator, and the National Co-Chair of the Research Group. He has been the President or the Co-President of international conferences or special sessions and conferences for international journals. He has authored or co-authored (more than 180) numerous book chapters, journal articles, and communications. He participates in several research networks, projects, and associations.

JOÃO MANUEL R. S. TAVARES graduated in mechanical engineering from the Universidade do Porto, Portugal, in 1992. He received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Universidade do Porto, in 1995 and 2001, respectively, and the Habilitation degree in mechanical engineering, in 2015. He is currently a Senior Researcher with the Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial and an Associate Professor with the Department of Mechanical Engineering, Faculdade de Engenharia da Universidade do Porto. He is the co-editor of more than 40 books, and the co-author of more than 35 book chapters and 600 articles in international and national journals and conferences. He holds three international patents and two national patents. He has been a Committee Member for several international and national journals and conferences. He is the co-founder and the co-editor of the book series *Lecture Notes in Computational Vision and Biomechanics* (Springer). He has been a (Co-)Supervisor for several M.Sc. and Ph.D. theses and a Supervisor for several Postdoctoral projects. He has participated in many scientific projects as a Researcher and as a Scientific Coordinator. His main research interests include computational vision, medical imaging, computational mechanics, scientific visualization, human–computer interaction, and new product development. He is the Founder and the Editor-in-Chief of the *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* (Taylor & Francis) and the Co-Founder and the Co-Chair of the International Conference Series, including CompIMAGE, ECCOMAS Vip IMAGE, ICCEBS, and BioDental. More information can be found at www.fe.up.pt/ tavares.