Int. J. Nonlinear Anal. Appl. 14 (2023) 3, 47–62
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.32871.4335

Using Hadoop to analyze big data for multiple purposes: An applied study according to the Map-Reduce model

Shereen Sh. Ahmed[a,*], Delveen L. Abd Al-Nabi[b]

[a]Department of Computer Science, Faculty of Science, Zakho University, Duhok, Kurdistan Region, Iraq

[b]Department of Economic, College of Economic and Administration, Duhok University, Duhok, Kurdistan Region, Iraq

(Communicated by Nallappan Gunasekaran)

Abstract

The volume and diversity of data in the world are unprecedented in human history. It is growing at an unprecedented rate. Internet and social media technologies as they permeate every stage of our lives and even our mobile phones, people have become a source of data even in their daily activities. So, a new concept emerged: "Big Data". Big data is produced with high volume, speed, structured diversity, and semi-structured and unstructured data. Many industrial areas release big data by creating new data or digitizing existing data models so that organizations can gain a competitive advantage. In order to extract economic value from big data, it should be processed with advanced analytical methods. This research aims to examine the use of Hadoop in analyzing big data according to the Map-Reduce model. and distributed file systems such as Processing, PIG, Mahout, NoSQL, and Cassandra, and the study concluded that advanced analytical methods protect the privacy of personal information, and through them, security gaps can also be filled, and the phenomenon of big data was discussed in terms of its components and resources, and it was emphasized on the advantages of big data in the areas of application.

Keywords: Hadoop, Big data, Map-Reduce, Big data analytic, organizations
2020 MSC: 94A16

1 Introduction

As a result of the great developments in the field of information and communication technologies in recent years, labor, information, which takes its place as a new production factor as well as capital and natural resources, It has become the most important resource element in providing competitive advantage. of organizations Innovation is an indispensable requirement in achieving sustainable competitive advantage. of innovation knowledge shapes its nature. Knowledge is the achievement of certain ends or a certain understanding. transform the data into useful formats for managers as a result of a transformation and analysis process [21]. In other words, in the decision-making process of managers the information they use is transformed into a useful and meaningful form by subjecting the data to information processing processes. the value that helps the success of the decisions made, and the data become meaningful and useful information. refers to the informational raw materials that need to be processed before they arrive [21]. In this context, making strategic decisions with a data-based approach It is of great importance in maintaining the sustainability of innovation-based competitive advantage is doing.

*Corresponding author
Email addresses: shereen.ahmed@uoz.edu.krd (Shereen Sh. Ahmed), delveen.luqman@uod.ac (Delveen L. Abd Al-Nabi)

Received: November 2022 *Accepted:* January 2023

Information and communication that organizations see only as a support provider in business processes for years technology solutions are taking place in many areas of our lives today and people are living and began to change the way they work. End users of information, data using multiple devices These devices record an ever-increasing number of events. personal computers In recent days, we are experiencing the transition from the Internet of Things (IoT) to the Internet of Things (IoT). RFID (Radio Frequency Identification) and sensor (sensor) technologies are becoming increasingly common. Coming from sensors and the Internet of Things, on websites, social media and mobile platforms. The data obtained by combining the data produced and the data within the organizations. Stacks have revealed the concept of "big data" [21]. In other words, big data with the increase in activities in digital environments, huge amounts of data reaching wider audiences stands for sets. CCTV (Closed Circuit Television) cameras, GPS (Global Positioning) System) and data recorded via sensor networks, increased communication through the use of digital texts platforms, photography and blog posts, large amounts of data potentially available for analysis [31].

The proliferation of smart devices used with the development of technology, the realization of the devices with the help of sensors. With the production of real-time event records (log), increasing mobility and internet access, social the fact that networks are a part of our lives day by day increases the diversity and speed of the data that surrounds us. and increasing its volume. This situation means that big data is received, stored and processed at the same speed brings with it problems. At this point, big data analytics is a structured enterprise selecting appropriate data from the unstructured data world such as video, audio, text file, provides access to information through its storage and processing. to social media applications, the use of sensors that collect data, and smartphones. Dependency has intensified the amount of data transmitted across networks. This transmitted data is generally unstructured. and are of the type obtained from different sources in different formats. Large and diverse Relational databases are insufficient to handle this type of data in their configurations, It supports the storage of unstructured data and has distributed parallel processing capability requires the use of systems [48].

While it is accepted that important benefits are provided in many areas with big data analytics, ethical regulations to mechanisms to be developed to ensure the privacy and security of users as well as other issues related to big data [28]. Because the potential to analyze big data, breach of privacy and Restriction of personal freedom areas brings with it concerns [30].

In this study; respectively, the concept of big data, its components and resources are discussed. The advantages of data in application areas are emphasized, big data analytics processes and Map-Reduce (Map-Reduce) running on the distributed file system that is the basis for big data analytics. Reduce computing model and working principles of Hadoop software architecture are examined, big data the issues to be considered in the field of security have been revealed and the security measures have been evaluated.

2 Big data

The concept of big data was first introduced by Michael Cox and David Ellsworth in 1997. At the IEEE Imaging Conference (Proceedings of the 8th Conference on Visualization), "Application- It is used in the article "Controlled Demand Paging for Out-of-core Visualization". In the same study, because the data sets are very large and you can save the computer system's memory, disks and even external disks.

It was mentioned that the problem was even filled in, and this problem was called the "Big Data Problem" [10]. Later, Francis X. Diebold, "Big Data Dynamic Factor Models for In his study titled "Macroeconomic Measurement and Forecasting", big data has been studied in physics, biology and which have to be faced in many fields of science, including social sciences, and refers to itself as a "phenomenon" that should be benefited from [14].

From this point of view, data is called "the raw material of our age". This fact, naturally Google, It has been known from the very beginning by giant IT companies such as Amazon, Twitter and Facebook, and even this issue lies at the basis of the founding philosophy of these companies [1].

Big data, which cannot be processed using traditional database techniques, it is a new concept that describes heterogeneous data in different volumes and consists of various digital contents. consists of [16]:

1. Structured data: Structured data, modeling, input as input, storage, querying, It refers to all data types that are easy to process and visualize. In general, certain types and sizes are presented in predefined fields, relational databases or tables. can be managed. In this data type, which has a solid structure, high performance of processes obtaining useful information because it does not require skills or parallel techniques easier than data types.
2. Semi-structured data: Semi-structured or self-describing data is a structured data. Although it reflects its type, it does not contain only a rigid model in its essence. in other words Semi-structured data includes models in

which structurality is defined, as well as specific elements and different types of data. miscellaneous, such as labels and signs used to define a hierarchical representation of fields It also includes meta models. XML is one of the best-known examples of semi-structured data. (Extensible Markup Language) and JSON (JavaScript Object Notation) programming languages takes.

3. Unstructured data: Unstructured data is presented and stored in other than a defined format record types. Usually in free formats such as books, articles, documents, e-mails consist of texts and media files such as images, audio and video. This type of data hard to present in a rigid way, such as NoSQL (Not only SQL) in data processing processes gave rise to new mechanisms.

2.1 Big data components

There are 3 main components (3V) that characterize the big data phenomenon: variety, velocity and volume [22, 26, 30, 32]. In addition to the 3V that defines big data in some sources, reality (veracity) and value 5V is mentioned instead of 3V by including its components [11, 13, 16, 17, 34, 49].

2.1.1 Variation

Big data can be produced in a wide range of any type and format, and these mixed data types There is no standard set or rules between them. Data is structured, semi-structured and unstructured. It occurs in three types [16, 17, 32, 35]. another one In other words, diversity refers to the structural heterogeneity in a data set. This heterogeneous structure 95% of it is unstructured data [17]. Thus, most of the data produced today is of unstructured type, and Facebook, Twitter and video It is fed from a variety of sources such as its content. Structured data can be easily kept in databases and can be labeled. From this point of view, big data has the same definition as structured data.

It has no format or length. The row and column information used for structured data is specific, storing unstructured data with relational database systems in an order, it is quite difficult to analyze. Because unstructured data is row and row in a relational database, cannot be stored in columns. Semi-structured data is a relational database such as an unstructured data type. Although it does not have a specific structure that can be placed in the tables, the separation of data or certain a data type that can be labeled to be put in a queue. Labeling, similar data together allows grouping. For example, a call center conversation records the name of the customer, the conversation time, speaking time and the subject of the complaint [32]. Call center as can be seen from the example, the speech recording data can be grouped, but the subject of the complaint is the data in the group cannot be placed in relational databases because it contains unstructured data.

2.1.2 Speed

Data is in constant motion. In this context, the analysis of data flow is one of the important issues for data scientists. Started to become one [11]. Big data generation rate is very high and this speed is increasing day by day. Speed for all business processes, not just big data is an important factor. From this point of view, the processes that will process and analyze the data are also important for big data should be at the same rate as production. As an example of how fast big data is produced, 2.7 billion clicks of likes and comments per day on Facebook [19], 350 thousand tweets per minute and 500 million tweets per day (Twitter Usage Statistics, n.y.), processing 50 billion messages per day on WhatsApp and the world More than 200 billion e-mails are sent and received daily across the country can be given. Apart from social media, it will set an example for the speed of data production. There are many workspaces available. For example, a jet plane can consume 10 terabytes of data every 30 minutes it flies. It collects by means of sensors [44]. Similarly Formula 1 car In the race, 20 gigabytes of data are produced by 150 sensors on a car [18]. As another example of the production speed of big data, at CERN (Conseil Européen pour la Recherche Nucléaire - European Organization for Nuclear Research).

In the “Hadron Collider” experiment, 1 petabyte of data per second is obtained through sensors. It can be given that it was produced [35].

2.1.3 Volume

One of the big data problems is the volume dimension. Because storing and accessing data requires innovative tools. For example, a doctor’s note about his patient is several kilobytes raw image files produced for the examination can be stored as a text file. megabytes, the results of more advanced diagnostic tools such as magnetic resonance are several gigabytes. can happen. Considering that this volume will increase as the number of additional tests performed

in the hospital, terabyte and even petabyte levels will have to be dealt with. If the patient's past data If it is necessary to use it in analyzes, it will be inevitable for the data volume to reach the exabyte level [11]. Big data will no longer fit into existing databases to volumetric levels much higher than terabytes and petabytes that cannot be processed with data analysis techniques reached. Smartphones, which have become a part and indispensable part of our lives, are IP (Internet Protocol) Many hardware, such as remotely controllable devices and smart meter systems.

They transfer the data they produce through various applications. Therefore, the produced, stored and There is an exponential increase in the amount of transmitted data. IDC (International Data), a research organization Corporation, in a study called "Digital Universe Study", data to be reached in 2020 The amount of data will be 44 times what it was in 2009 and the annual data volume will reach 35 zettabytes estimated [35]. As a multinational company in the field of information technologies According to a report published by CSC (Computer Sciences Corporation) operating in 2020, It is predicted that the data volume to be obtained in 2018 will increase by 4.300% compared to today [40]. Today, even in medium-sized organizations, 1 terabyte volume data can be produced in a very short time and this data is highly diverse by many sources can be created. According to IBM (International Business Machines), as of 2014, While about 90% of the data was produced only in the last 2 years, 2.5 exabytes of data every day production takes place (What Is Big Data?, n.d.). total produced in human history up to 2003 While the amount of data was 5 exabytes, today the same amount of data can be obtained in just 2 days are produced.

2.1.4 Reality

Reality shows how accurate or reliable big data is. data, business should be reliable enough to be used in decisions. The high diversity of big data, analysis It complicates the process of ensuring the quality and reliability of the data obtained [8]. Data quality with reality, which is a very important dimension of big data is being evaluated. Because reliable models can only be produced with high quality data.

Unfortunately, most of the data is affected or can be at a certain noise level. Other In other words, anomalies such as the presence of outliers or missing values in the data can be detected [11]. These anomalies have some data source specific unreliability. associated with it. For example, customer sentiment reflected in social media, although although it contains valuable information, it is uncertain in itself because it requires personal judgment can accommodate. Therefore, the need to deal with uncertain and uncertain data, tools and analysis methods developed for data mining and management of uncertain data reflects another aspect of big data that needs to be addressed with the use of [17].

At this point, the accuracy and validity of the collected data is extremely important carries. Large amounts of data that are not correct or valid are both the basis for analysis not only will it work, but it may also lead to misinterpretations [16].

As the big data collected can cause statistical errors and misinterpretations, it is valuable. In order to obtain information, the authenticity of the data is of critical importance. Establishment of reality; get confirming the data obtained, reducing the noise level, revealing the relationship sequence, and it consists of the stages of determining the level of deception [21].

Therefore, data inconsistency The uncertainty that will occur in the data due to various factors such as lack of [13]. At this stage, the goals and objectives some purification controls should be established in obtaining the data. However, very large amount of data and heterogeneous data source, ensuring the integrity and value of the inputs Strict use of both in the organization of the collected data and in its cross-checking should be subject to rules [16].

2.1.5 Value

Big data, high volume, speed, which requires specific technology and analytical methods to transform into value and information assets characterized by diversity. Based on this definition, the value of big data is emerged with its transformation into an understanding that can create economic value for organizations and society.

Will emerge (Mauro, Greco and Grimaldi, 2016, p. 131). Big data, often "low density value" means. In other words, the data received in the original form is usually low compared to its volume in the first place. has a value. However, a high value can be obtained by analyzing such large volumes of data [17]. The value of big data lies in interrelated individuals and by making connections between pieces of data about groups or about the structure of the information itself obtained from accessible patterns [5].

After the production and analysis stages of big data, the organization adds added value to business processes [33]. Instant access to big data's decision-making processes It is very effective in terms of value component, being immediately accessible in making the right decision, is important. To be able to benefit from the cost advantage

that big data will add to business processes, It will make big data valuable because it is more than the cost that organizations will bear for [34].

2.2 Big data sources

Today, there are more data sources than yesterday. Smartphones, tablet computers, sensors, medical equipment, web traffic logs, interactions in social networks and pharmacy, meteorology, many sources, such as scientific research that offer solutions in areas such as simulation, use big data [39]. However, the increasing heterogeneity of the web environment, web pages in different media (e.g. text, image and video), genres (e.g. encyclopedia, news, blogs) and topics (e.g. entertainment, sports, technology) [2].

A large number of data sources are effective in increasing the diversity of big data. One of these sources while some of them can be completely new data sources, some data sources are decomposition of existing data, In other words, it emerges as a result of transferring existing resources to digital media. Many The industrial field falls under the umbrella of new data generation and digitization of existing data, and every one of them constitutes a separate big data source. Industries that are growing big data are as follows can be listed [35]:

1. Transport, logistics, retail, utility and telecommunications: Transport, logistics, GPS transceivers used in retail, utility and telecommunications industrial areas, Via RFID tag readers, smart meters and sensors in phones Data is being collected at an ever-increasing rate. This collected data is used to optimize operations, instantaneously. Recognizing emerging business opportunities and developing organizational business intelligence can be used for operation.
2. Healthcare: The healthcare industry is rapidly developing electronic medical imaging and moves towards benefiting from reporting. Electronic medical imaging and reporting data, short-term public health monitoring and long-term epidemics needed for use in research.
3. Government: Many government agencies, census, energy use, budget reports, law enforcement It digitizes public reports such as election results and election results and makes them accessible to the public offers. This type of data is held by government agencies and regional communities and is widely available. It is data that can be used in business and management applications operating in a wide range of applications. This data the vast majority are freely accessible on the web, while some are available for a fee can be obtained in return.
4. Entertainment media: such as books, newspapers, magazines, television, radio, movies, cinema, music and games. The entertainment industry, which serves in many areas, has increased its popularity in the last 5 years, such as digital recording, production and showed a transition towards distribution. Today, people and societies in entertainment media a wide range of data is collected observing their behavior.
5. Life sciences: Low-cost gene as an example of data generation in the life sciences industry count can be given. Gene that can be performed at less than $1,000 census, analysis to investigate genetic diversity and to determine potential treatment efficacy constitutes tens of terabytes of data that can be
6. Video display: In the video display industry, IP from subtitled television technology Progress has been made towards based television cameras and recording systems. IP based analysis for the development of new technological camera data, security and services are collected for.

3 Application areas of big data

Big data helps researchers find answers to their questions, individual behavior and community. It provides convenience in estimating their trends [21]. However, economic and from commercial activities to public administration, from national security to scientific research big data is used. One of the important goals underlying big data applications some are improving consumer experiences, reducing costs, better marketing strategies and increasing the efficiency of existing processes. Also, today's data Establishing security due to the occurrence of breach events is also the purpose of use of big data started to be among them. Major application areas of big data include banking, communication, media and entertainment industry, health services, education, manufacturing, government services, insurance, retail and trade, transport, energy sector and analysis of self-measurement data takes.

According to the statistics of the research organization Statista, as of 2016, big data and analytics in the overall market share, banking was the application area that generated the most revenue with 13.1%. Banking, 11.9%, batch production, 8.4%, process-type production, 7.6%, government services and 7.4%, respectively followed by professional services. In the same year, the total of big data in all application areas its market value has reached 130.1 billion USD (Statista, 2016). another researc institution IDC, this total revenue value obtained in 2016, has a compound annual growth rate of 11.7%. predicts that it will reach levels of more than US$ 203 billion by 2020 (Press, 2017).

3.1 Big data applications in banking

Earn more profit from past datasets with big data analytics in the field of banking compared to yesterday is being done. Historical data shows cash movements, predictable disasters, robberies and customer guides the understanding of their behavior. With the use of big data, banks, money can see the details of their movements, predict disasters and theft and can better understand consumer behavior [33].

In the international arena, banks, analysis of customer behavior, cross-selling of products, compliance with regulations from the power of big data in many areas such as management, risk management, coping with financial crimes started to benefit [43].

3.2 Big data applications in communication, media and entertainment industries

Cinema production, television broadcasting, news, communication and gaming thanks to big data organizations have begun to face new business models. In this situation, customers access to content offered from anywhere and from any device originates. Today, we embrace creativity, multi-channel promotion and payment methodologies.

- There is increasing pressure to develop. These methodologies are based on customer It is developed depending on the understanding of the media usage trends and activities of the profile. In addition, as the interest of the computer-based customer profile moves towards media tools, customer demands are increased. The chances of adapting the content increase accordingly. Thus, media and entertainment organizations, big data can use its resources to ensure more customer participation [33].
- Media organizations in the world, to gain competitive advantage in the globalizing media market and to It uses big data to better tailor content for viewers. Big data It provides the opportunity to predict what the audience wants before they even know it, and present the content accordingly. Big data, such as bringing customized internet search results with feedback permeating our daily lives, helping organizations overcome bottlenecks, customer behavior It helps them understand and improve their organizational performance [45].
- Social media, which is a means of communication and socialization, is taking its place in human life day by day is enlarging. The increasing use of smartphones and the expansion of high-speed mobile networks, It reveals the culture of instantaneous uploading of data produced by people to web pages takes out. To illustrate the magnitude of this phenomenon, the number of photos uploaded on Facebook reaching 4,000 per second, 243,000 per minute, 14.58 million per hour and 350 million per day can be given as an example (Aslam, 2018). Social media is among the most used areas of big data measurement of customer satisfaction. customers about products and services.
- Organizations value customer feedback in order to keep a close eye on their thoughts [25]. Unstructured data such as text files community data to big data obtained by transferring them to dynamic networks that detect trends it is called. Likes on a web page that reflects consumers' thoughts about a product.
- The data obtained from the buttons, comments shared on Twitter, constitutes an example for community data [18]. In this sense, big data is available via social media. It also guides the marketing activities to be carried out. For example, online complaint by contacting a customer who has shared a bad experience with a company's products on their site offers can be submitted or the customer's feedback is evaluated and deficiencies in the products can be made.

3.3 Big data applications in healthcare

The amount of data produced in the field of health services continues to increase day by day with diseases Health records of individuals struggling to survive are among the important sources of big data takes. Big data provides an overview of the patterns and trends of certain diseases and offers the opportunity for early diagnosis [33]. Simultaneously reducing costs Analyzing big data effectively in order to increase the quality of health services required. In order to improve the quality of health services, patient-centered service delivery, infectious early detection of diseases, monitoring the quality of hospitals and improving treatment methods Big data is used in many fields such as [3]. E.g In 2005, 30% of doctors and hospitals in the US used electronic medical records, By the end of 2011, this rate has increased, with approximately 75% of hospitals and doctors 45% use electronic medical records. In addition, approximately 45% of hospitals in the United States HIEs (Health Information Exchanges) where local and regional health information is shared as of 2013 platforms and will continue to exist in the information sharing platforms of hospitals in the future [23].

With the access to electronic medical imaging and reporting data in health services, people Special treatment methods can be developed according to their genetic characteristics and health history. Also A state institution providing public service in health services, region, province, district, etc. instantly to see the distribution of diseases, doctors and hospitals at different levels, the efficiency of the service to be given to the citizen will increase its level. As another example of big data usage areas in healthcare live case demonstration applications can be given. Live case display, to support patient-oriented and on-patient teaching of the images of the surgeries performed, and to be shared on various social media platforms in order to raise awareness of the society.

3.4 Big data applications in education

Big data is a quality resource for both teachers and students by many educational organizations can be used to ensure the continuity of the educational environment. For example, what do students time they log in, the web pages they browse, how long they spend on the pages, and from big data in revealing the general pattern of events, such as their activities over time can be utilized. The number of students, demographic structure, demands and education subjects.

Teachers' educational activities are also measured and regulated in dimensions [33]. In addition, online technology emerged with the combination of internet technology and traditional education. Big data in (online) education is the restructuring of education structure, scope, technologies and methods plays a central role in its restructuring [47].

3.5 Big data applications in manufacturing

To support decision-making processes in the fields of production and sourcing and to compete in this context information from the geographic, graphical, textual and temporal elements of big data estimation models are used [33]. In addition, smart manufacturing Emerging applications, such as process and product lifecycle management, together with big data began to come to life. Active preventive maintenance in smart manufacturing systems, big data implemented through analytics. With the support of big data in the production area, production devices alarms, device event Many real-time device data such as device records and device status notifications can be collected [46].

3.6 Big data applications in government services

As computer-based data continues to increase and reaches unpredictable dimensions, information is stored, management, processing, security and regulation become more difficult. smartphone apps, Due to the increase in sensors and cloud computing solutions, governments' data generation and archiving rates are also rising. Public institutions and organizations that collect, research and analyze big data can benefit from unstructured data through new tools [33]. In government services, petabytes of data are produced every day. This data is real-time analysis, to governments to improve the quality of education, reduce the unemployment rate, retirement opportunity provision, delivery of aid to all those in need, live streaming data on traffic such as controlling traffic density and improving mobile ambulance services will help in providing value-added services to its citizens in many fields [3]. In addition, the services offered on the e-government portal with big data analytics efficiency and productivity can be increased. At this point, big data means that the services provided to the citizens are fast and It plays a key role in the development of smart cities by ensuring that they are reliable [49].

3.7 Big data applications in insurance

With the use of big data in the insurance field, better price adjustment and more robust by establishing customer relations, the profitability and performance of insurance organizations can be increased [33]. The region where the customers live, age, insurance status, gender, customer are the most important big data elements used in predicting profitability [15].

3.8 Big data applications in retailing and commerce

Big data flow in retail can be visualized in five dimensions: Customers (with each customer) relevant detailed data), products (data related to product features and levels), time (real-time data), location (geo and destination data), and channels (data from all channels) [6]. Key benefits of using big data in retail between the correct display of stocks, timely analysis, shopping patterns optimization of staff employment and customer and ensuring continuity in their relations [33].

3.9 Big data applications in transportation

Public institutions and organizations, controlling traffic, planning the best transportation route, smart transportation to develop systems, predict traffic conditions and manage congestion that may occur. can benefit from the data. In the private sector, consolidation and consolidation of shipments thanks to big data by optimizing transport movements, advances in technological solutions, increase in revenues and competitive advantage can be achieved. In order to save fuel and time individually, Big data can be used to plan the appropriate transportation route. Similarly, tour the use of big data in regulations can also provide ease of access [33]. In the field of traffic management, smart transportation where information and communication technologies are widely applied with the development of GPS transceivers, CCTV systems, detectors, microchips, mobile such as road conditions, vehicle and driver behaviors collected via mobile phones and other portable devices. traffic data constitutes big data. Fast and dynamic developed with the use of this data models can provide better simulation capabilities for intelligent transportation systems [50].

3.10 Big data applications in the energy sector

Providing better resource and workforce management with big data, It helps to detect and quickly review before it happens. E.g Instead of the old meters that collect information once a day, smart meters that collect information every 15 minutes more effective control over customers' consumption information and energy infrastructure [33]. In the energy sector, sensors, cloud computing technologies .With the use of wireless and network communication in applications, large amounts of data are increasingly obtained. Big data is changing energy production and consumption patterns. Energy big data is only smart It does not contain meter reading data, but also other other data such as weather data and geographic information system. It also encompasses a large amount of data from sources. For example, energy production and consumption data, geographic information system data and weather data (temperature, atmospheric pressure, humidity, cloud cover, wind speed and wind direction) integration, location of renewable energy generation devices It can guide the determination of power generation and energy efficiency [51].

3.11 Big data applications in self-measurement

Data produced by individuals measuring personal activity and behavior, self-measurement It is called self-quantification data. For example, people's movements, exercises monitor and analyze the data by transferring the data obtained from it to the smart phone application. wristbands that provide self-generating measurement data. Individuals in psychology Despite their behavior, there are other things they want to do in line with their fixed tendencies. E.g a person may want to minimize energy consumption by purchasing an energy saving light bulb. How is this time Even though it consumes less energy, it can keep the lamp on for a long time. The self-measurement data is at this point, It bridges the gap between psychology and behavior. social policies such as psychology, marketing, and public policy. In these areas, personal fixed-prone and indirect behavior data can be used [18].

4 Big data analytic

Each project begins to achieve a specific goal and sub-goals are determined in line with this goal. Organizations use appropriate data, available data as raw material to achieve project goals and objectives. It has to reach from data sets that are too large and complex to be processed by information systems.

After obtaining the appropriate data, the structured, semi-structured and unstructured data collected as a whole will associate, process and reach information in the stage of making strategic decisions in projects data analytics will be needed. With the technological developments, it is now possible to supplement the structured data.

As a result, semi-structured and unstructured data types have also started to be used [17]. Big data analytics is the process of processing very large and diverse records containing different types of content. It is the use of analytical and parallel techniques developed on behalf of big data analytics at this point tools are rapidly changing and very complex tools that are difficult to process using traditional database techniques. With the analysis of the structural, semi-structured and unstructured data as a whole, valuable data can be obtained from the data aims to obtain information [16]. In other words, big data analytics is the use of large data sets to obtain information that will guide the decision-making process. It is a technique used to analyze the data [17].

An example of big data analytics application through an organization working in the retail industry can be given. Increasing sales with big data analytics or marketing, pricing, stock, advertising and a bridge between organizational subunits responsible for areas such as customer relations and business ideology task can be established. In line with the objectives, sub-goals are determined and within this framework, Data is collected from inside and outside the

organization. Some of this collected data is available for a fee while some may be in the category of freely available data available to the public. Primarily event records, stock movements, recorded transactions, customer Structural data such as information, fee information and supplier information will be collected. In addition to the structured data call center and help desk incident records, e-mail or other customer feedback obtained through communication channels, stock traffic generated by sensors and unstructured data such as parking space usage records will also need to be collected. important here, the point is that the data that will not affect the result of the analysis phase in reaching the goals and targets is It is the elimination of the data in order to reduce the performance loss in accordance with the reality component. Data In the next step of the collection phase, the unstructured, from customer feelings to geopolitical issues governments, research companies, social networks, etc. by many other sources external data produced will be collected. Data from outside the organization as well as inside the organization those should be cleared of fake or invalid data before entering the analysis process. In analysis phase on the other hand, unstructured big data elements are used in order to obtain maximum value from big data automatically or purposefully developed application, reporting and it will need to be combined with inquiry tools [39].

From the retailer example as it is understood, non-structured records such as web records, customer feedback and call center records such as purchasing, customer and supplier records kept in relational databases of the organization with data. Analyzing the data in structural format together, in line with the determined goals and objectives, will ensure that the highest value is obtained from the data. The main questions that need to be answered in big data analytics are: Big data size and as the diversity of data increases, how will the problems to be encountered in data analytics be dealt with? all data should it be stored? Should all data be analyzed? Which big data elements really how should it be decided that it is important? How to get the best advantage of big data should it be used? These unanswered questions also pose great challenges in the analysis phase of big data brings with it. Big data consists of structured, semi-structured and unstructured data types.

For this, advanced capabilities in big data analytics are required. However, on the data the type of analysis to be performed also depends on the results to be obtained. In the analysis phase, either all big data elements are combined, or it is determined which big data element is relevant to the result to be achieved [25]. At this point, organizations are working on big data analytics four key capabilities to expand their business intelligence systems and analytics initiatives. Should take into account. These capabilities include advanced analytics, imaging and research, turning intuition into action and obtaining the right mix of information. Advanced analytics, the most important capability of big data analytics and it serves to reveal closed patterns. As new data types arrive, asset analytics, network there is a need for the use of new algorithms such as analysis, text content analysis and real-time scoring will be heard. In order to increase the accuracy and reliability of big data, users can use additional data may want to add resources or increase data volume. This highlights the importance of scalability.

In this context, in big data analytics, these new algorithms use text, image and should be used to make sense of video content. Imaging and research capabilities It can help organizations find answers to business-related questions. new data types high volume, revealing different patterns, highlighting key points This increases the need for new imaging modalities such as temperature maps that provide at this point, Tools such as Tableau Software and Datameer are interactive, iterative, research-based and provides visual data exploration. The third of big data analytics act on intuitions capable of with conversion, decisions are produced by both manual and automatic processes. big data big data analytics processes that automate decision making and forecasting models and business models that set boundaries when judgmental evaluation is needed. reveals the need for technological infrastructures that use the rules. Big data analytics as its fourth capability, analytical tools mix the right information for business purposes [29].

Researchers and analysts are using the high volume, variety and speed of unstructured data today manages with tools based on relational database management system used for structured data they face difficulties [26]. Compared to structured data, unstructured data the fact that the size and growth rate is too high means that unstructured data can be used in databases.

The need for high storage space and energy consumption are among the main difficulties. However, in relational databases such as structured data, unstructured data can be arranged in a certain order. Lack of interoperability due to lack of storage is another challenge. Query languages running on and on data storage systems to overcome these challenges The logic has been changed. In this context, relational databases in big data analytics instead of distributed file systems, relational database query languages such as SQL (Structured Query Language) instead, open-source software specific to big data started to be used. big data Among the software used in analytics, mainly Map-Reduce technology and HDFS (Hadoop Hadoop software architecture using Distributed File System is coming. Hadoop currently includes Yahoo, Amazon, Facebook, Linkedin, Twitter, IBM and Adobe. It is used in big data analytics in many leading IT companies. with these big data solutions are not limited to Hadoop. Although open-source software although Hadoop is the most well-known among them, big data analytics can deal with a single solution

method. It has a complexity that cannot be solved [29]. MPP in big data management (Massively Parallel Processing) [29], NoSQL [16, 29], HBase, HCatalog, PIG, Mahout, Cassandra and In-memory [16] Other technologies such as MPP databases are generally unshared scaling architectures, each node maintains a subset of the MPP database, and thus providing high autonomy regarding the execution of parallel queries [22]. Hive is a data tool that allows you to retrieve and manage large amounts of distributed data warehouse tool. Hcatalog, on the other hand, is used in the Hive tool, regardless of whether the data is structured or not.

It is a table management system that allows it to be stored in a format. Hbase is Hadoop's database. A distributed data model based on the BigTable file system used by Google base management system. Increasing performance in PIG, Hadoop and Map-Reduce technologies It is a software platform that uses the "PIG Latin" programming language that provides Mahout, Map-Reduce Filtering distributed datasets developed over Hadoop using technology, It is a software that contains a set of algorithms for classifying and clustering. NoSQL, non-relational Querying and calling semi-structured and unstructured data on all databases represents an enabling concept. Cassandra, column-driven, developed by Facebook It is a NoSQL database and supports Map-Reduce technology and is especially useful for large amounts of records provides the ability to facilitate data access. In in-memory processes, instead of hard disks RAM memory of hardware is used, big data is long-term despite archiving problems are not stored, and this facilitates instant access to the right information [16]. The said technologies used in big data analytics, especially IBM, Kognitio and as enterprise solutions by many product providers, including ParAccel & SAND presented [37].

4.1 Hadoop software architecture

Distributed file system primarily used by Google in the creation of big data analytics and the distributed computing model Map-Reduce played an important role. with file system by Google in the related article published in 2003, how to store data with distributed file system [20], in the article published in 2004 model, how data can be queried and compiled over this distributed file system [12]. Inspired by these two articles, the large Early versions of Hadoop, a software architecture capable of handling data distributed Doug Developed by Cutting and Mike Carafella. The name "Hadoop" comes from Doug who developed the architecture. It comes from the name of Cutting's son's toy elephant [39]. Hadoop after developing it, Doug started working at Apache. Hadoop, still maintained by Apache It is presented as an open-source Map-Reduce framework written in the Java programming language [7]. Despite the growing data volume and speed, Hadoop is the only scaling, such as being stored in a domain, only with no license fee required offers relatively inexpensive routes by spending money on equipment [22]. In other words, the Map-Reduce model with Hadoop has been added to the Google file system. Distributed file system that can be built with similarly large and inexpensive server clusters is operated on Hadoop architecture has two main components, HDFS and Map-Reduce. on Hadoop the operations to be performed are typically written as a job and given to the HDFS server cluster.

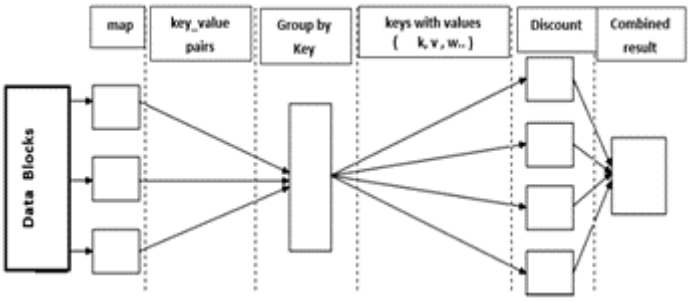
HDFS works in cluster computing logic. It also means parallel processing architecture. In inbound cluster processing, servers with processing nodes are kept in racks. Each a cabinet can contain 8 to 64 nodes. Servers (nodes) among themselves Gigabit Ethernet while connecting with the network interface, the cabinets also communicate with each other via switching devices is in the state. Files are divided into blocks, usually 64 megabytes, called "chunks" [27]. Block size and number of times blocks are copied to nodes can be determined by users. In Hadoop architecture, if a node fails by duplicating a solid copy of the content on another node, the data being securely stored are kept. Also, block sizes in HDFS the default value is usually greater than 64 megabytes to reduce the number of disks needed. Determined [25]. However, it keeps copies of a block. Nodes should be located in different cabinets against cabinet deterioration. At this point, the data block what their size will be and how many times the blocks will be copied, It should be determined by the users according to the level at which it can be accessed. HDFS is a combination of ordinary servers as in parallel processing architecture. It allows very large files to be stored on a single virtual disk that will be created (by clustering) offers. The HDFS server cluster, which constitutes the file system of the Hadoop architecture, is generally the main It consists of two different types of nodes, namely the master node and the worker node. [25].

Being a master node, job tracker, task tracker and name node It performs three different functions. Data entries received from the user application with the job tracker Accordingly, Map-Reduce tasks are distributed to worker nodes. task follower, within the cluster It offers nodes the ability to import Map-Reduce tasks from the job tracker. If the name node It holds the index information (metadata) of each file block in the HDFS file system. When the user application wants to access the file, it contacts the name nodes. Mother having more than one node is very important for business continuity [39].

Each worker node has its own data node and task tracker. Also each worker node stores the data distributed on it, lists the blocks it owns. Periodically reports to the name node and copies the data to other nodes in the cluster executing the process. If the worker node is represented as the address by the name node, the user contacting the

application. Unlike master nodes, worker nodes are many and come together high-throughput analysis capable of analyzing data volumes in the hundreds of terabytes or even petabytes when they arrive reveals its power [39].

Map-Reduce, complex big data developed by Google using the smash-conquer method based on the fact that the problems are first fragmented and then processed in parallel on many servers. It is a programming framework [39]. Map-Reduce, as the name suggests, Map (Map) and Reduce [4]. In this context, MapReduce dividing large problems into small, manageable sub-problems, as shown in Figure 1, then key-mapped data blocks that distribute them across multiple servers called nodes grouped and analyzed at each sub-node, then brought them together again to the desired result. It is a reductive software framework. In the peer stage, the master node in the distributed system model It takes its inputs, divides them into sub-jobs, and distributes them to worker nodes. Worker nodes are in job tracker control fulfill the sub-tasks given to them with the task follower returns and the resulting data pairs (key, value) to local filesystems that can be accessed by the next Downgrade stage. places it. In the reduction phase, the master node receives the results from the worker nodes and converts them to the key data. [4, 41, 27]. See Fig. 4.1.



Source: Leskovec et al., 2014, p. 25

For example, let's say we have a complex file with words on each line as input. Match-Reduce model can be used to find how many of each word. With the word counting example, the operation of the Match-Reduce model is simulated and a large amount of presented data is used. (eg at terabyte level) small amount of data (eg at megabyte level) as a result of analysis It is shown that it can be produced [4]. In the word counting example in Fig. 2. As shown, the Hadoop server cluster consists of 2 nodes and the file input is split into 2. In the match phase, each node matches within itself according to the keyword and evaluates the number of information. assigns as. Then, grouping is done according to the keys and the reduction stage is started. In the reduction stage, the values corresponding to the keys in each group are assigned to their own values as per the given job description. are added together and the results are reduced. Thus, which word can be escaped with Match-Reduce? It is concluded that there are many [41].

Source: Siddesh et al., 2014, p. 238

5 The security dimension of big data

Although big data is a tremendous innovation for many industries and decision makers, many users It also poses great security risks. These risks are caused by big data tools. storing, managing, analyzing various data collected from all accessible sources, arises from its visualization and sharing. certain behavioral data big data generating, especially internet users, due to the discovery and consolidation of individuals become vulnerable to the exposure of sensitive information. In other words, big In data analytics, it may be possible to collect more data than necessary, which results in causes many security and privacy violations [16]. In confidential data big data due to increased security risk and adoption of cloud computing technologies Ensuring security in the environment emerges as an important issue. Transaction today power and storage areas are easily available and cloud computing technologies are adoption is

increasing the data volume gradually. This makes the data accessible to the outside world. Big data security measures are needed to prevent data from falling into the wrong hands [8].

While big data offers new opportunities to a wide range of society, some of these opportunities are data collection. unpredictable at this stage. For example, when blood samples were collected from people forty years ago, blood DNA testing was not mentioned as a potential use of the samples. with this It is a fact that should be taken into account that data can be abused. in slums Practices such as not giving credit / not providing insurance to living people are examples of abuse [9]. Collecting personal data in the last days of the information revolution, use and analysis is essential. However, at this point, individuals do not know which data about themselves They do not know exactly what is collected and shared with third parties. Confidentiality, transparency and If important values such as identity information cannot be protected, innovation and advantage with big data There is a risk of losing these values for the sake of obtaining them. The concept of privacy in the age of big data, It needs to be better understood in order to manage the flow of personal data. with this It is also important that the data with confidentiality value remains confidential during sharing.

In addition, big data should be transparent and identity information for individuals to feel safe [36]. At this point, privacy and personal data regulations, both internationally and nationally, in order to guarantee the confidentiality of has been brought. The first comprehensive international agreement on the protection of personal data, No. 108 adopted by the European Council on 28 January 1981, "Automated Data Protection Agreement on the Protection of Persons Against Processing (Personal Data Protection Agency (KVKK), 2018, p. 11th). In article 7 of the said contract, "automatic to protect personal data saved in files, accidental or unauthorized destruction of them or accidental loss or unauthorized acquisition, alteration or appropriate security measures are taken against the distribution of it" (Official Gazette, 2016a). In addition, in our country, in the Constitution of the Republic of Turkey, "everyone, private has the right to demand respect for his life and family life. Private and family life its privacy is inviolable" (Constitution of the Republic of Turkey, 1982, Article 20).

The "Personal Data Protection Law" numbered 6698, which was prepared on the basis of the aforementioned article, defines the concepts of personal data, sensitive personal data and the processing of personal data and explains the terms of processing the data. According to the Personal Data Protection Law, personal data, It means "any information relating to an identified or identifiable natural person". special qualification Personal data refers to the "race, ethnic origin, political opinion, philosophical belief, religion, sect or other beliefs, clothing, association, foundation or union membership, health, sexual life, criminal conviction and security measures, and biometric and genetic data. Processing of personal data means "wholly or partially automatic or any data processing of personal data". to be obtained, recorded by non-automatic means, provided that it is part of the registration system, storage, preservation, modification, rearrangement, disclosure, transfer, such as taking over, making available, classifying or preventing its use means any operation performed on the data". Both personal data and private qualified personal data, other than the procedures and principles stipulated in this Law and other laws, It cannot be processed without consent (Official Gazette, 2016b).

By ensuring the security of big data, unauthorized corporate and individual data with confidentiality value intended to be protected from access. Confidential data, accessible from public sources It may be data that does not contain confidentiality value belonging to companies, organizations or individuals. For example, a transaction data belonging to the bank customer, RFID tags used in the organizational supply chain, company goods and movement of resources many other, such as website browsing logs and mobile phone usage. Personal or organization-specific data that can be produced in the field has the value of confidential data [18].

Traditional data security approaches cannot be applied to big data management. At this point, there are new security approaches that organizations with data warehouses must address. Prudent organizations seek to protect the value they derive from big data investments understand the risks posed by big data and directly address those risks. Firstly big data assets within the organization should be defined, information assets that need to be protected should be classified and the risks on these assets should be evaluated used in big data,. Most of the tools are used by large IT companies to analyze specific data for a specific purpose was developed for. At this point, depending on the data types analyzed, there was no need for security and since big data tools are developed with open source code, it can also be used for general use security functions may be overlooked. Although the tools used are different, weak access controls (identification, authentication, authorization and access control), insecure communication infrastructure, poor user interface, security, such as the absence of encryption and the presence of vulnerabilities that could lead to penetration attacks vulnerabilities. However, there is a completely hopeless situation in the field of big data security.

As a matter of fact, the development of security functions in many big data tools continues is doing. In this context, new companies offering security-oriented big data solutions such as SQRRL it is seen that it has started to enter the

market [24].

Keeping data in one place makes it a target for attackers. There fore big data stores should be kept under control effectively at this point. Identity to ensure authentication, a cryptographically secure communication infrastructure must be established. Controls should be implemented and constantly monitored, especially through access privileges. Possible prevent unauthorized and abusive use in the event of a personnel or authorization change controls need to be updated. However, metadata, packet information, and event Other security procedures that allow network traffic information to be monitored and analyzed, such as logs should be created [38].

With the concept of data privacy, it is emphasized that information about the private life of individuals is kept confidential. In this direction, the personal data voluntarily uploaded to social media platforms throughout the society awareness about its value, potential and risk has begun to emerge. Privacy on social media the subject can be dealt with in two classes. The first of these is the social situation with insufficient protection by individuals themselves. are the privacy problems that may occur as a result of uploading data to the media environment. The other issue that deals with privacy is the large scale created by other users' media. covers data issues. What makes this problem particularly bad is that the injured person unaware of the process of uploading the relevant data to the Internet and therefore any protective [42]. to social media platforms an exemplary case that can be given regarding the privacy of uploaded personal data is the 2016 It happened during the presidential election campaigns of the year.

The data environment around us is called data exhaust. Waste data is used for many purposes. However, when combined with other data elements, it can turn into a form of value. For example, in mobile phone use, a person spreads some data about his daily activities. like shopping when health services are used, or when communicating with other people. When passed, data broadcasting is done passively. Limited or zero mana at first An information with confidentiality value emerges from the combination of this type of data, which can express [18]. Jeff Jonas, chief scientist at IBM, On her blog "Using Transparency as a Mask": "Twenty years Unlike before, people today love their relationships, your relationships, your photos, your large volumes of usable data by describing your photos and opinions about you they produce. The more data, the more understanding and predictive capabilities it offers. Different maybe a secret date of yours will be exposed or the most intimate You will want to stop talking to your friend. However, you know that all this information is confidential. you were thinking." [32]. Scott Charney from Microsoft in 2012 During his speech at the RSA Conference, he provided a good example of data privacy: "If a friend of mine takes a picture of me playing volleyball and shares it with other friends and if one of them uploads this picture to Facebook, my insurance company can find this picture and send it to me. can be used against it." In his speech, Charney said that health insurance companies are using this and similar data emphasizes that it can increase insurance premiums [9]. With the privacy of the data Another relevant example can be given from the practice of live case display in healthcare. Alive in the informed consent of the patient who will participate in the case presentation, confidentiality and privacy boundaries should be clearly defined. Informed consent, "to be applied to the patient himself to be sufficiently informed to approve or reject any medical procedure, It refers to the process of thinking about information and making a decision based on free choice". (Informed Consent Guide, 2013). Retail the use of big data in predictive analytics in the industry brings ethical and privacy issues together brings. Making new laws and regulations in order to protect the privacy of consumer data such as state interventions may be needed [6]. As can be seen, big data has many advantages, especially in social media, health services and marketing. In addition to its potential to be beneficial in the field, it also has potential negative effects on the privacy of private life brings with it. At this point, personal data obtained by combining different sources There is a general fear in society about its inappropriate use [35]. Personal The new information obtained by combining the information with large large data sets collected from outside, may reveal a fact that has a degree of confidentiality and that he does not want anyone to know [25]. Another dimension of the privacy problem in big data is It is what companies that control almost all of big data do with the information they get from the data.

This concern determines what data companies can collect about us and can be overcome by regulatory interventions that allow what to do with the data. with this Another important issue that needs to be emphasized is that other people's data also affects us. This issue is important both for our acquaintances to see this data and for companies that carry out big data analytics. it is related to obtaining this data and should be addressed in a social sense [42].

As a result, managing privacy effectively in big data is a technical and sociological (techno-social) issue. is the problem. The benefits of big data can only be found by addressing privacy within these two perspectives. will be understood [35].

6 Conclusion and recommendations

In this study, big data phenomenon, application areas of big data, big data analytics and big data security dimensions have been investigated. In this context, the benefits to be obtained with big data applications big data analytics processes are examined on the basis of Hadoop and Map-Reduce, security that can endanger the privacy of personal information, as well as the advantages of data problems were highlighted and the precautions to be taken regarding big data security were technical and social. size was evaluated. The increase in the size, diversity and production speed of data in the world has led to the concept of big data and technologies have emerged. In this context, social media, marketing, health big data analytics in many fields such as banking, insurance and public services. benefits began to be perceived. Taking full advantage of big data structured data for In addition, unstructured data, which constitutes 95% of big data, is also included in data analytics processes. necessity has been raised. Like SQL, which is currently used for structured data, relational database management systems are inadequate for analyzing unstructured data types. Remains. Advanced analytics running on distributed file system for analysis of unstructured data methods are used. Among the software used in big data analytics, MapReduce and Hadoop software architecture based on HDFS. Thanks to Hadoop, Google the Map-Reduce model on the file system is compatible with multiple and inexpensive server clusters.

It can be run on distributed file systems that can be created. Big data software architectures protect data privacy and security because they are developed with open source code. may be insufficient to ensure safety. A generally accepted software in big data analytics and the lack of enforcement discipline is what organizations need to take to ensure data privacy. increasing measures. While it is a known fact that big data has the potential to be beneficial to humanity in many areas, The fact that personal data may endanger the privacy is also a security issue that should be addressed. emerges as a subject. Data that does not make sense at first can be combined with different data. may come and reveal an information with confidentiality value. At this point, big data managing societies of those who have and are successful in transforming big data into information It is an undeniable fact that they will achieve their ability. While we humans create data, big data The fact that the majority of them are in the hands of only a few giant search engines and social media companies comes to mind. brings a new phenomenon, the "big data threat". From big data for the benefit of humanity both technically and sociologically in terms of violation of the privacy of personal information. security measures must be taken. Technically, in big data tools, primarily identification, Access controls such as authentication, authorization and recording of accesses should be implemented and encryption technologies should be used in data communication. In the sociological dimension, first of all, national Big data phenomenon should be handled in all its aspects in international and international platforms, and legal arrangements should be made. In addition, with the vehicles with sensors, smart phones and internet It should be acted with the awareness that data is created almost every minute through activities and in this context, Big data awareness should be brought to the society with its pros and cons.

References

[1] T. Acarer, *Opportunities and problems offered to software companies by the pandemic process*, Eurasia Proc. Educ. Soc. Sci. **22** (2021), 18–25.

[2] S. Achsas, *Improving relational aggregated search from big data sources using deep learning*, Intelligent Systems and Computer Vision (ISCV), IEEE, 2017, pp. 1–6.

[3] J. Archenaa and E.M. Anita, *A survey of big data analytics in healthcare and government*, Procedia Comput. Sci. **50** (2015), 408–413.

[4] K. Bakshi, *Considerations for big data: Architecture and approach*, IEEE Aerospace Conf., IEEE, 2012, pp. 1–7.

[5] D. Boyd and K. Crawford, *Six provocations for big data*, A decade in internet time: Symposium on the dynamics of the internet and society, 2011.

[6] E.T. Bradlow, M. Gangwar, P. Kopalle, and S. Voleti, *The role of big data and predictive analytics in retailing*, J. Retail. **93** (2017), no. 1, 79–95.

[7] P. Chandarana and M Vijayalakshmi, *Big data analytics frameworks*, Int. Conf. Circuit. Syst. Commun. Inf. Technol. Appl. (CSCITA), IEEE, 2014, pp. 430–434.

[8] S. Chandra, S. Ray, and R.T. Goswami, *Big data security: survey on frameworks and algorithms*, IEEE 7th Int. Adv. Comput. Conf. (IACC), IEEE, 2017, pp. 48–54.

[9] S. Charney, *Trustworthy computing next*, Tech. report, Microsoft, 2012.

[10] M. Cox and D. Ellsworth, *Application-controlled demand paging for out-of-core visualization*, Proc. Visualization'97 (Cat. No. 97CB36155), IEEE, 1997, pp. 235–244.

[11] B. Cyganek, M. Graña, B. Krawczyk, A. Kasprzak, P. Porwik, K. Walkowiak, and M. Woźniak, *A survey of big data issues in electronic health record analysis*, Appl. Artific. Intell. **30** (2016), no. 6, 497–520.

[12] J. Dean and S. Ghemawat, *Mapreduce: Simplified data processing on large clusters*, Commun. ACM **51** (2004), no. 1, 107–113.

[13] J. Debattista, C. Lange, S. Scerri, and S. Auer, *Linked'big'data: towards a manifold increase in big data value and veracity*, IEEE/ACM 2nd Int. Symp. Big Data Comput. (BDC), IEEE, 2015, pp. 92–98.

[14] F.X. Diebold, *Big data dynamic factor models for macroeconomic measurement and forecasting*, Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society," (edited by M. Dewatripont, LP Hansen and S. Turnovsky), 2003, pp. 115–122.

[15] K. Fang, Y. Jiang, and M. Song, *Customer profitability forecasting using big data analytics: A case study of the insurance industry*, Comput. Ind. Engin. **101** (2016), 554–564.

[16] Y. Gahi, M. Guennoun, and H.T. Mouftah, *Big data analytics: Security and privacy challenges*, IEEE Symp. Comput. Commun. (ISCC), IEEE, 2016, pp. 952–957.

[17] A. Gandomi and M. Haider, *Beyond the hype: Big data concepts, methods, and analytics*, Int. J. Inf. Manag. **35** (2015), no. 2, 137–144.

[18] G. George, M.R. Haas, and A. Pentland, *Big data and management*, Acad. Manag. J. **57** (2014), no. 2, 321–326.

[19] B. Gerhardt, K. Griffin, and R. Klemann, *Unlocking value in the fragmented world of big data analytics*, Tech. report, Cisco Internet Business Solutions Group, 2012.

[20] S. Ghemawat, H. Gobioff, and S.-T. Leung, *The google file system*, Proc. Nineteenth ACM Symp. Oper. Syst. Principles, ACM, 2003, pp. 29–43.

[21] P.B. Goes, *Editor's comments: Big data and IS research*, MIS Quart. **38** (2014), no. 3.

[22] N. Golov and L. Rönnbäck, *Big data normalization for massively parallel processing databases*, Comput. Standards Interfac. **54** (2017), 86–93.

[23] P. Groves, B. Kayyali, D. Knott, and S.V. Kuiken, *The'big data'revolution in healthcare: Accelerating value and innovation*, Res. Brief **7** (2016), 1–11.

[24] O. Hamami, *Big data security: Understanding the risks*, Bus. Intell. J. **19** (2014), no. 2, 20–26.

[25] A. Katal, M. Wazid, and R.H. Goudar, *Big data: issues, challenges, tools and good practices*, Sixth Int. Conf. Contempo. Comput. (IC3), IEEE, 2013, pp. 404–409.

[26] E.W. Kuiler, *From big data to knowledge: An ontological approach to big data analytics*, Rev. Policy Res. **31** (2014), no. 4, 311–318.

[27] J. Leskovec, A. Rajaraman, and J.D. Ullman, *Mining of massive data sets*, Cambridge university press, 2014.

[28] S. Manca, L. Caviglione, and J. Raffaghelli, *Big data for social media learning analytics: potentials and challenges*, J. e-Learn. Knowledge Soc. **12** (2016), no. 2.

[29] B. Mantha, *Five guiding principles for realizing the promise of big data*, Bus. Intell. J. **19** (2014), no. 1, 8–11.

[30] C.L. McNeely and J.O. Hahm, *The big (data) bang: Policy, prospects, and challenges*, Rev. Policy Res. **31** (2014), no. 4, 304–310.

[31] S.J. Miah, H.Q. Vu, J. Gammack, and M. McGrath, *A big data analytics method for tourist behaviour analysis*, Inf. Manag. **54** (2017), no. 6, 771–785.

[32] M. Minelli, M. Chambers, and A. Dhiraj, *Big data, big analytics: Emerging business intelligence and analytic trends for today's businesses*, vol. 578, John Wiley & Sons, 2013.

[33] K. Naik and A. Joshi, *Role of big data in various sectors*, Int. Conf. I-SMAC (IoT in Social, Mobile, Analytics

and Cloud)(I-SMAC), IEEE, 2017, pp. 117–122.

[34] R. Narasimhan and T. Bhuvaneshwari, *Big data-a brief study*, Int. J. Sci. Eng. Res. **5** (2014), no. 9, 350–353.

[35] F.J. Ohlhorst, *Big data analytics: Turning big data into big money*, vol. 65, John Wiley & Sons, 2012.

[36] Neil M Richards and Jonathan H King, *Big data ethics*, Wake For. Law Rev **49** (2014), no. 393, e432.

[37] Philip Russom et al., *Big data analytics*, TDWI best practices report, fourth quarter **19** (2011), no. 4, 1–34.

[38] S. Sagiroglu and D. Sinanc, *Big data: A review*, Int. Conf. Collaborat. Technol. Syst. (CTS), IEEE, 2013, pp. 42–47.

[39] S.A. Schneider, *'big data:'big challenge, big opportunity*, 2012.

[40] K. Setty and R. Bakhshi, *What is big data and what does it have to do with it audit*, ISACA J. **3** (2013), no. 14, 1–3.

[41] G.M. Siddesh, S. Hiriyannaiah, and K.G. Srinivasa, *Driving big data with hadoop technologies*, Handbook of Research on Cloud Infrastructures for Big Data Analytics, IGI Global, 2014, pp. 232–262.

[42] M. Smith, C. Szongott, B. Henne, and G. Von Voigt, *Big data privacy issues in public social media*, 6th IEEE Int. Conf. Digital Ecosyst. Technol.(DEST), IEEE, June 2012, pp. 1–6.

[43] U. Srivastava and S. Gopalkrishnan, *Impact of big data analytics on banking sector: Learning for indian banks*, Procedia Comput. Sci. **50** (2015), 643–652.

[44] H. Sun and P. Heller, *Oracle information architecture: An architect's guide to big data*, Oracle, Redwood Shores, 2012.

[45] J.J. Tang and K.E. Karim, *Big data in business analytics: Implications for the audit profession*, The CPA journal **87** (2017), no. 6, 34–39.

[46] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas, and A.V. Vasilakos, *A manufacturing big data solution for active preventive maintenance*, IEEE Trans. Ind. Inf. **13** (2017), no. 4, 2039–2047.

[47] S. Yu, D. Yang, and X. Feng, *A big data analysis method for online education*, 10th Int. Conf. Intell. Comput. Technol.d Automat. (ICICTA), IEEE, October 2017, pp. 291–294.

[48] R. Zafar, E. Yafi, M.F. Zuhairi, and H. Dao, *Big data: The NoSQL and RDBMS review*, Int. Conf. Inf. Commun. Technol. (ICICTM), IEEE, May 2016, pp. 120–126.

[49] N.Z. Zainal, H. Hussin, and M.N.M. Nazri, *Big data initiatives by governments–issues and challenges: A review*, 6th Int. Conf. Inf. Commun. technol. Muslim World (ICT4M), IEEE, November 2016, pp. 304–309.

[50] J. Zeyu, Y. Shuiping, Z. Mingduan, C. Yongqiang, and L. Yi, *Model study for intelligent transportation system with big data*, Procedia Comput. Sci. **107** (2017), 418–426.

[51] K. Zhou, C. Fu, and S. Yang, *Big data driven smart energy management: From big data to big insights*, Renew. Sustain. Energy Rev. **56** (2016), 215–225.