Computer Assisted Methods in Engineering and Science
31(3): 313–349, 2024, doi: 10.24423/cames.2024.906

This article belongs to the *Special Issue on AI-Based Future Intelligent Networks and Communications Security* edited by Dr. S. Kumar, Dr. G. Mapp, Dr. A. Bansal and Dr. K. Cengiz

A Study of Big Data in Cloud Computing

Imran KHAN

Department of CSE, Harcourt Butler Technical University, Kanpur, U.P., India;
e-mail: imran.k@hbtu.ac.in

Over the last two decades, the size and amount of data has increased enormously, which has changed traditional methods of data management and introduced two new technological terms: big data and cloud computing. Addressing big data, characterized by massive volume, high velocity and variety, is quite challenging as it requires large computational infrastructure to store, process and analyze it. A reliable technique to carry out sophisticated and enormous data processing has emerged in the form of cloud computing because it eliminates the need to manage advanced hardware and software, and offers various services to users. Presently, big data and cloud computing are gaining significant interest among academia as well as in industrial research. In this review, we introduce various characteristics, applications and challenges of big data and cloud computing. We provide a brief overview of different platforms that are available to handle big data, including their critical analysis based on different parameters. We also discuss the correlation between big data and cloud computing. We focus on the life cycle of big data and its vital analysis applications in various fields and domains At the end, we present the open research issues that still need to be addressed and give some pointers to future scholars in the fields of big data and cloud computing.

Keywords: big data, cloud computing, distributed computing, data mining, Hadoop.

1. Introduction

Over the last two decades, the amount of electronic data has increased enormously due to the data generated from various sources. According to a report by the International Data Corporation (IDC), the approximate volume of electronic data in the world was around 1.8 zettabyte (ZB), i.e., 10^{21} bytes, in 2011 alone [12]. This volume increases day by day. With the latest advancement in information technology (IT), organizations are producing and storing more and more data. For instance, on YouTube, videos are often uploaded at a rate of about 72 hours every minute [14]. Organizing and analyzing the produced data in order to gain insights is still a challenge but is crucial for competitive advan-

tage [46]. Analytical solutions that can analyze a variety of organizational data (structured and unstructured) are highly desirable, as they help organizations gain insights from their privately owned data as well as the data available on the web [115]. This ability to analyze cross-relate private and publicly available data, such as tweets, blogs and data from other social networking sites, enables organizations to better understand their customers' needs and preferences, as well as predict their future needs. This opens a new paradigm for various researches and is popularly termed big data.

Big data refers to rapidly growing data that cannot be efficiently managed by traditional database platforms in terms of storage, processing, and analysis [9, 10, 144]. Also, current tools and systems are insufficient to address the distributed computational needs and exploit the large variety of data [11, 145]. Despite the popularity of big data and analytics, managing and gaining insights from big data is a complex, time-consuming and costly task. Big data brings new opportunities and offers considerable value to organizations that are willing to accept it [47], because it provides various options to gain in-depth understanding of internal as well as external functionalities of an organization and helps in finding various hidden facts, although realization of these added values remains a considerable challenge [46]. Manyika *et al.* [9], based on McKinsey Business Technology Office research, focused on the value of big data in the U.S. medical industry, suggesting that if big data could be utilized effectively and efficiently the industry may surpass USD 300 billion in value annually. Organizations willing to leverage big data and analytics often acquire high-cost software licenses, employs high-processing machines and infrastructure, and hire expert analysts who understand the business scenario and can built solution for understanding customer needs, behavior, and future demands for new products [48]. Despite the higher costs involved, industries, government agencies, and academic institutions have recently shown increasing interest in big data research and applications [13], as evidenced by initiatives by the governments of the USA [49, 146], UK [50, 51] and Russia [52], and academic initiatives such as big data initiative by MIT [53], and efforts by companies like Intel [54].

Cloud computing has revolutionized the information technology (IT) industry by improving resource utilization and enabling organizations to pay only for the resources they use on a time basis. Organizations of all sizes are using cloud computing to reduce IT capital, operational expenditure and other resources.

Although cloud computing differs significantly in terms of technologies and implementation, it generally offers three basic services: infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS) and software-as-a-service (SaaS) [55, 56]. In the paradigm of the Internet of Things (IoT), sensors and various instruments generate web logs that are stored and processed in cloud storages. Cloud computing provides several benefits in which the most common is offering resources

on pay-as-you-use basis. This allows small and mid-scale organizations to use IT resources without purchasing them, thereby reducing the costs of infrastructure and maintenance. With the rapid growth of data, the need to store and manage it arises and creates problem for traditional systems to change their previous techniques, so cloud computing could be a viable solution as it offers improved availability and elasticity. The goal of this research is to look into the current state of big data and cloud computing. We discuss cloud computing as well as the definition, characteristics, classification, and life cycles of big data. We also investigate the tools and techniques used in big data management and analysis.

2. Defining big data and cloud computing

2.1. Big data

Since 2011, a new paradigm of research known as "big data" has emerged [57], significantly changing our reality and revolutionizing large-scale data researches. The worldwide big data market is expected to more than double in size by 2027 from its estimated value in 2018. The software industry will have a 45 percent market share by 2027, making it the largest segment of the big data market [143] (Fig. 1). So far, diverse concepts, theories and definitions on big data have been proposed by academia and industrial researches. Among these rapidly evolving definitions, few have created confusion, some have focused on *what big data is* part, while others emphasize *what it does* part from various perspectives. Although there have been extensive discussions in the past on defining big data [16, 17] in addition to formulating an appropriate definition, researchers also focus on how to refine its value. The initial thought that occurs when trying to elucidate the concept of big data is size or volume.

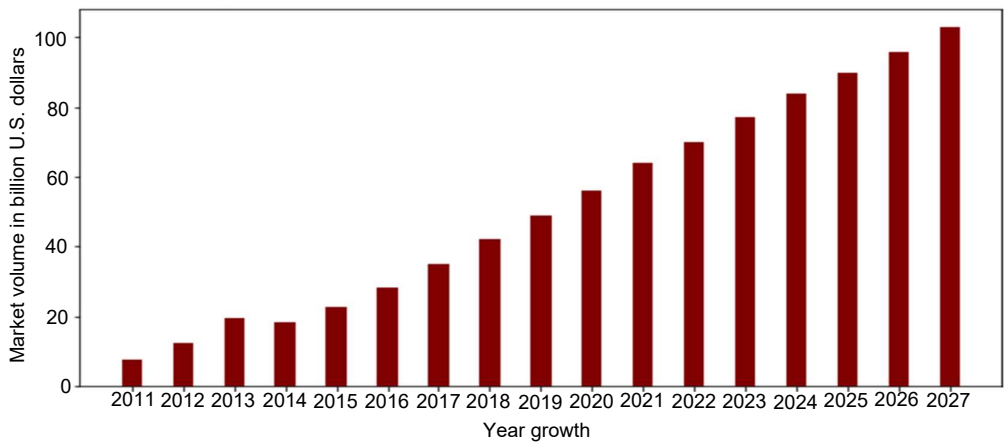

Fig. 1. Big data market forecast.

Volume is not the only parameter to exemplify big data, other imperative characteristics have also emerged. For instance, the three characteristics: volume, velocity and variety (3 V's) are considered as the dominant challenges associated with big data and its management [1]. The 3 V's framework has emerged as the most popular way for describing big data [2, 3]. Many researches on big data have used this 3 V's framework, as defined by Gartner, Inc., in the following way: "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" [4].

Similarly, big data is defined by National Institute of Standards and Technology (NIST) as "consisting of extensive datasets – primarily in the characteristics of volume, variety, velocity, and/or variability – that require a scalable architecture for efficient storage, manipulation, and analysis" [58]. We describe the 3 V's of big data, i.e., volume, variety and velocity below.

The first V, i.e., volume, refers to the magnitude of data generated. The magnitudes of big data are reported in petabytes and exabytes. For example, authors in [5] reported that around one million photographs per second are processed by Facebook at peak. However, the concept of volume in the context of big data is relative and changes over time and depending on factors such as data type. Therefore, what may be considered large data now may not meet the prerequisite for big data in the future, as storage capacities of devices increase, allowing much bigger dataset to be stored easily and efficiently. In addition to this, the type of data also plays a significant role when defining how 'big' big data is. Data management technologies for two datasets of same size can be greatly different, for instance, handling tabular data versus audio or image data. Therefore, it is highly impractical to establish a universal threshold for big data volume, as it varies from organization to organization and their types of data generation.

The second V, i.e., velocity, refers to the speed at which data is generated and the rate at which it is analyzed. With the upsurge in electronic data-generating devices such as smartphones, sensors, server logs, etc., has led to a high volume of data generated at a very high rate. Even conventional retailers are generating data at a very high speed. For example, Wal-Mart processes more than one million transactions per hour [6].

The third V, i.e., variety, refers to the different types of data generated. With technological advancements organizations generate a variety of data, including structured, semi-structured and unstructured data. Structured data specifies data that can be stored in spreadsheets or relational databases. Semi-structured data includes headers. tags and other markers for structuring web data. Unstructured data, on the other hand, does not have any structuring (predefined data models) and can include text, images, audio and video.

In addition to 3 V's, few definitions embody two additional V's, expanding to the concept of 5 V's in defining big data [7]. These additional two V's include veracity and value (Fig. 2). Veracity refers to the data consistency and data trustworthiness. It ensures data integrity and authenticity across various sources.

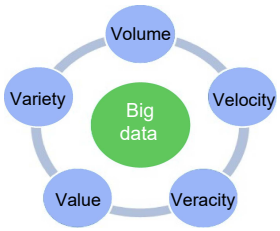

Fig. 2. Big data characteristics.

Value refers to the benefits that the big data will bring to events and processes. This characteristic is eminently linked to volume and variety, as organizations assess how much and what type of data to store and for how long.

It is important to note that no universal benchmarks exist for the three characteristics of big data (volume, velocity or variety). Their defining limit may vary based on location, industry sector and size, evolving over the time. Also, these three dimensions are dependent on each other, i.e., changes in one may impact the others. However, a tipping point for the 3 V's exist for every organization, beyond which traditional mechanisms or techniques for managing data become inadequate, prompting the search for more robust and effective mechanisms to handle big data [8].

The inability of traditional IT mechanisms used by organizations to handle large datasets, along with different opinions from research scholars, data analysts, enterprise experts and technical practitioners, has brought some more definitions on big data. In general, big data could be defined as the datasets that cannot be captured, managed or processed by conventional hardware/software tools in a reasonable timeframe [12, 15].

Based on the above definition, in May 2017 McKinsey & Company reported big data as a cutting edge concept for innovation, competition, and productivity [9].

2.2. Cloud computing

Due to the increase in the complexity of the computational world, a new paradigm offering efficient and economical solutions for large-scale and complex computing tasks has emerged known as "cloud computing". Cloud computing has revolutionized modern information and communication technology (ICT) and its services, moving away from traditional means to provide a new and more robust

architecture for performing complex and large-scale processing tasks. It is a fast growing and popular technology for IT industries and other businesses that use the IT services. Cloud computing delivers reliable software and hardware over the Internet from remote data centers, accessible through mobile devices [38]. Cloud computing offers so many services (such as SaaS, PaaS, IaaS) and various advantages like parallel processing, virtualized resource pooling and scalable data integration and storage in a cost-efficient way. Cloud computing not only provides the benefit of optimized resource utilization for individuals and enterprises but it also reduces infrastructure setup and maintenance cost through efficient management [37]. With all the above said, cloud computing popularity has led to a significant increase in the size of data handled by applications leveraging its benefits.

Cloud computing has emerged as a powerful architecture and service model that allows ubiquitous, on-demand and convenient access to various computing resources (e.g., storage, applications, software and hardware) [42]. Many organizations are increasingly adopting cloud computing to store, process and analyze large datasets [39]. Many scientific applications are migrating to the cloud for extensive experiments due to lack of available local server computing capability and the cost efficiency and efficient management offered by cloud platforms for large-scale experimental data [40]. Cloud service providers have also started to integrate parallel data processing frameworks to assist users so that more applications and programs can be deployed on the cloud [41]. Cloud computing offers organizations a total cost ownership model so that they can focus on core business rather than on various issues like infrastructure, availability and management of resources [43]. Also, utility model of cloud computing offers a wide range of infrastructure, computation and storage services, which has highly attracted the attention of scientific community worldwide [44]. The service model of cloud typically consists of three main services: PaaS, SaaS, and IaaS.

PaaS provides customers with platform computing to develop and run applications. Examples include Salesforce.com, Force platform, Microsoft Azure and Google's Apps Engine. SaaS offers subscription-based licensing on different software applications for end-users, which can be accessed through the internet [45]. It includes Gmail, Google Docs, Online Payroll and Salesforce.com. IaaS offers virtualized computing resources and hardware equipment on demand basis. Examples include Flexiscale and Amazon's EC2. Table 1 lists some of the popular service providers.

Numerous organizations are using cloud services to deploy and analyze their data. Cloud computing offers mainly three deployment models: private, public and hybrid.

Private cloud: The private cloud model is set up on a private network and is run either by the company itself or by a third party. The fact that it is

Table 1. Cloud platforms.

Platform	Computing architecture	Service	Load balancing	Fault tolerance	Storage	Programming framework
Amazon Web Services [125]	Elastic Compute Cloud	IAAS	Round robin load balancing	Automatic alert on failover and resync back to last known state	Simple Storage Service (S3)	Amazon MapReduce framework
GoGrid [126]	Data center architecture	IAAS	F5 load balancing, round robin and SSL least connect algorithms	Highly scalable and reliable with file-level backup	Two-step storage: first connection setup with private network and then using transfer protocols to transfer	Java, Python, Ruby
Flexiscale [127]	Data center architecture	IAAS	Automatic equalization of server load	Fully self-serviced	Persistent storage using SAN	C, C++, C#, Java, PHP, Perl, Ruby
Google App Engine [128]	Google geo-distributed architecture	PAAS	Automatic load balancing	Automatic	Big Table distributed storage	MapReduce
Azure [129]	Microsoft data center architecture	PAAS	Built-in hardware load balancing	Containers are used	SQL server data services	Microsoft.Net
RightScale [130]	Multi-server clusters	PAAS	High availability proxy load balancing	Basic, intermediate and advanced failover architecture	Open storage model	Ruby, Amazon Simple Query Service
Eucalyptus [131]	Multiple clusters	IAAS	Simple load balancing	Separate clusters within the main cluster reduce the chance of failure	Walrus	Axis2, Axis2c, Java

implemented and managed privately makes it suitable for businesses that need complete control over data security and confidentiality. In this paradigm, data sharing only takes place inside divisions of the same company.

Public cloud: The public cloud is deployed offsite and is available for the general public to develop and deploy their applications. It is managed by the service provider with the focus on quality of service in terms of privacy, security, availability, etc. It allows to use the resources at a lower cost.

Hybrid cloud: The hybrid cloud combines both private and public cloud. It allows analytical application to be developed in the private cloud and uses additional resources from the public cloud as needed. The decision to opt for a specific model is typically made by top management based on organizational requirements and budget. The fundamental technology used in cloud computing implementation is virtualization.

Virtualization [59] is a process that allows resource sharing and isolation of underlying hardware, so that the resource utilization, scalability and efficiency will increase. It abstracts logical resources from their physical counterparts. Many big data environments use virtualization as a basis technology which is required to access, store, process and analyze large scale datasets.

3. Cloud computing and big data

Cloud computing and big data move hand in hand. Big data utilizes cloud computing services to perform computationally intensive operations. Cloud computing aims to provide high computing power and extensive storage capacity resources to big data applications on-demand. The development of cloud computing enhances the capabilities of big data as it provides various computing and storage services to big data. Also, with the emergence of big data the cloud computing paradigm has also accelerated. Various technologies used by cloud computing like distributed computing and parallel computing have improved the efficiency of effectively managing and analyzing big data. Firstly, big data relies on cloud computing infrastructure for smooth and cost-effective functioning. Secondly, the target customers of big data and cloud computing are different. But with the increasing demand of both the technologies they intertwine with each other. With the increasing demand for data and analysis, big data has evolved, while cloud computing has evolved from the development of virtualization technologies. These technologies supplement each other. The development of cloud computing not only provides computation and processing facilities to various big data applications but it also operates on its service mode. Also, big data applications have expanded the service base of cloud computing. Therefore, both the domains provide enhancement to each other [19, 20].

Big data and cloud computing are closely related. Big data makes it possible to process dispersed queries using cost-effective computing resources. Cloud computing uses the Hadoop architecture for performing distributed data processing. Large datasets from the web are stored on distributed platform and processed through various programming models using parallel distributed algorithms [63–65].

Big data utilizes cloud services for distributed storage rather than relying on local server storage, thereby ensuring high and efficient storage. The big data technologies have greatly advanced due to rapid growth of cloud computing and virtualization technologies. Therefore, cloud computing provides computation and server as service facilities for big data due to which the two technologies are growing hand by hand (Fig. 3).

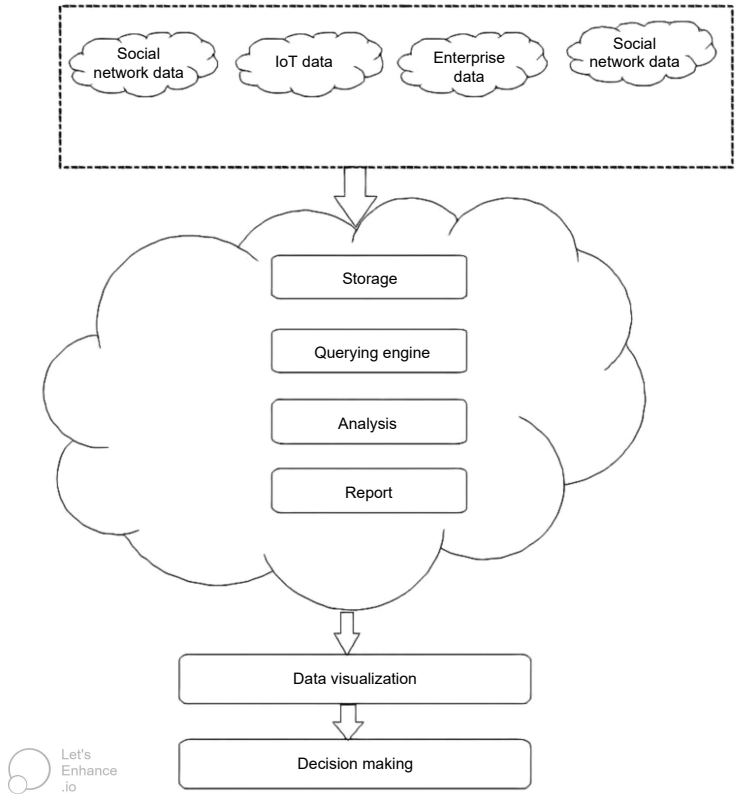

Fig. 3. Big data in cloud.

4. Life cycle of big data

The life cycle of big data is divided into four main stages, i.e., data generation or data sources, data storage, data analysis and the available tools and techniques

to handle big data (Fig. 4). In the following sections, we will study each stage in detail.

Fig. 4. Life cycle of big data.

4.1. Data sources

The first step in the life cycle of big data is data generation, and it includes identifying the sources from where big data is generated. As we all know, the internet is the biggest source of electronically generated data. It includes data from various sensors, internet forum posts, search entries, chat records, and many more. The data is closely related to users' daily routine life and has high value. Individually, such data may be useless, but if accommodated and analyzed using big data techniques, it provides valuable information, such as hobbies, purchasing behavior and it can predicts the mood of the user as well. So, identifying all these sources of big data from the internet is highly desirable. The following are the commonly acceptable sources of big data but make sure these are not the only sources of big data.

4.1.1. Enterprise data. Enterprises generate a huge amount of internal data that consists of online trading data and historical analysis data generated from internet sources. In addition, enterprise data includes production data, inventory data, finance and sales data, etc. Most of the data generated in the enterprises is managed by relational databases management system (RDBMS). The scenario has changed over the last decades; digital data on the internet have contributed a lot to the development of organizations and enhanced profitability. It is estimated that the volume of enterprise data in the world doubles every 18 months [9] which will generate a turnover of more than USD 450 billion through the internet, enterprise-to-enterprise and enterprise-to-customer interactions [21]. This increase in the volume of enterprise data requires more sophisticated solutions to handle and harvest the potential of large-scale data. For example, Amazon processes more than 500 000 queries from third-party sellers per day [1]. Akamai analyzes more than 75 million events per day for its marketing and advertising. The Walmart store chain processes more than 1 million transactions per hour a day and imports these records into their databases with a capacity of over 2.5 petabytes (PB) [22].

4.1.2. IoT data. One important source of big data is IoT, and many conversations on big data cannot be completed without mentioning IoT. The data

may come from passive, active or dynamic sources [66]. IoT data may include mobile data, web logs, sensors data, data form smart cities, such as transportation, medical care or public departments. It generates large-scale data. In IoT, a large number of data acquisition devices are distributively deployed, which may acquire data from various locations, or multimedia data like surveillance video data. The data generated from these devices are heterogeneous, and there is a high correlation between location of the data and its timestamp. In order to analyze data from these sources, both current and historical data will be equally required.

4.1.3. Bio-medical data. In the field of bio-medical informatics, due to the high throughput bio-measurement technologies, the size of data generated is very high, opening up a new paradigm of big data. For example, ProteomicsDB [67] covers almost 92% of human genes and has a data volume of about 5.17 terabytes (TB). Also, the new generation of gene sequencing technologies produces data so rapidly that billions of DNA sequences are produced daily at a low cost. Since fast speeds are required for gene sequencing, big data technologies are tailored to handle this. By developing smarter, more accurate, and efficient analytical solutions for bio-medicine applications can drive future development, helping in developing more sophisticated devices as well as producing more research and development of new drugs. The completion of the human genome project (HGP) and the advancement in sequencing technologies have led to a wide range of big data applications that support specialized investigations. The results of these analyses could be used for early diagnosis and personalized treatment of diseases.

4.1.4. Social network data. With the advent of social computing and the 'mushrooming' of social media devices, a new paradigm of big data have emerged known as social big data analysis. A typical internet user shares and consumes a large amount of digital data via popular online services such as Facebook, Twitter, YouTube, and Instagram, among others [68]. The data generated on these platforms is highly varied, as it contains various multimedia content like text, images, videos, etc. Such data contributes to 95% of all big data [8]. Due to its sheer volume and highly informative content, such data opens various possibilities in processing and analyzing it for personal [69, 70], commercial [71, 72] and societal purposes [73–75]. For example, the commercial use of such data includes more targeted advertising, matchmaking services, and many data-centric business models [77, 78]. As the data generated through these social networking platforms is so huge that it cannot be tackled using the previous traditional systems. So we need to develop more robust and efficient tools that can leverage social big data.

4.2. Big data storage

With enormous volumes of big data accumulating rapidly, compliant storage architecture needs to be implemented to efficiently store, retrieve, and manage data chunks, while also ensuring data accessibility and integrity. The growing demands for storage mechanisms requires infrastructures to be more robust in terms of storage space and simultaneously provide a more preeminent mechanism for querying, storing and analyzing data chunks. These requirements entail auxiliary equipment such as servers, data storage mechanisms used to store, manage, search, and process data with structured RDBMSs. With the advent of data growth due to technological advancements, data storage devices are becoming essential, and many global organizations resort to larger capacity devices to remain competitive.

Many researchers and organizations are working towards devising such competent mechanisms of storage. Persistent mechanisms for storing big data chunks are primarily classified into three bottom-up levels: 1) distributed file systems, 2) database systems, and 3) big data programming models. The first category serves as the baseline for higher-level applications. One of the most popular among them is Google file system (GFS) [23]. GFS uses low-cost commodity servers to help customers attain fault-resistant mechanisms and high-performance services. GFS is particularly suited for read-intensive large data applications rather than write-intensive ones. Despite its many benefits, the evolving mechanism of GFS a has some drawbacks in the form of a single point of failure and poor performance margins for small-scale data applications. In order to address these limitations, Colossus [24], the successor of GFS has been introduced. Numerous data intensive organizations have evolved customized solutions to meet their needs. For instance, HDFS and Kosmosfs are spin-offs of open-source codes of GFS. Microsoft's Cosmos [25] aids its scrutiny and advertisement business. Facebook utilizes Haystack [81] to store the vast number of small-sized photos generated through its processes.

4.2.1. Distributed file system. The biggest milestone in the development of big data processing is addressing the need for geographically distributed data storage systems across the globe. A distributed file system have emerged as a fine solution by dispersing data onto different geographically dispersed networks. It organizes various files and directory services from multiple servers into a global directory and provides access to data in a location-independent manner. All the files are accessible to the users through the global file directory structure.

4.2.2. Databases. Over the past three decades, the technological scenario has been witnessing manifold advancements in the database science. Varied ap-

TABLE 2. Different data stores for big data.

Database type	Database name	Advantages	Disadvantages
Key-value database	Dynamo	Highly fault tolerant, simple and easily scalable	Less efficient with complex data, foreign key need to be created
	Voldemort	High scalability, high fault tolerant, persistent hash table	No built-in support for multiple data centers
Column-oriented database	BigTable	Support semi-structured data, naturally indexed, highly scalable	Poor performance with the inter-connected data
	Cassandra	Configurable consistency level optimized for write queries	Reads are more disks intensive, does not support client conflict resolution
Document database	MongoDB	Simple and powerful model, easily scalable	Poor performance with the inter-connected data, only key and indexes are used for query processing
	SimpleDB	Easily scalable, no need for data administration, parallel query execution	Limits data size, some latency in data retrieval, no joins
	CouchDB	Uses very intuitive and well-designed HTTP-REST based interface, easily scalable	Complexity in querying, less support of JSON data
Graph database	Neo4j	Highly useful for connected data, easily represent semi-structured data	Allows only vertical scalability, data migration is highly complex
	AllegroGraph	High performance, partitioning with federation, support for secondary indexes	Less support for querying, single server data storage
	ArangoDB	Configurable consistency, memory efficient, support for complex queries	Horizontal scalability is difficult, no MapReduce support

plication requirements catering to datasets of multiple dimensions have been taken into consideration while developing these database technologies. Due to the limitations posed by traditional relational databases in dealing with massive datasets of big data, the advancements have gained pace. One of the most popular variants of these new advancements is the NoSQL data store. This database technology offers multiple key features in terms of flexible data models, ease of scalability, simple API sets, eventually consistent data and handling of large amounts of data with efficiency. NoSQL is the nucleus technology in big data realm. Further in this paper, we will examine the four major database types (Table 2): 1) key-value databases, 2) column-oriented databases, 3) document-oriented databases, and 4) graph databases.

4.2.3. Programming models. Big data models are hosted across several thousands of commercial application servers. In response to this, conventional parallel models like open multi-processing (OpenMP) and message passing interface (MPI) are inadequate to host large-scale parallel processing applications. Few ad-hoc solutions have been implemented in terms of enhancing programmatic models and the performance of NoSQL data stores as well as mitigating the performance drawbacks of typical traditional data storage models. Such advancements made these models the cornerstone technology for analyzing big data.

4.2.3.1. MapReduce. Another essential variant of the NoSQL technology stack is MapReduce. It is a powerful model for digitizing numerous clusters of commercial computers in order to attain the benefits of parallel processing and distribution. MapReduce is comprised of two computing functions – map and reduce, which can be programmed by core users. The map function generates intermediate key-value pairs from input key-value pairs. Once this is successfully completed, MapReduce brings together all the intermittent values related to the same key and transmits them to the reduce function, which further compresses the value set into coherent chunks of related set. Users can easily specify these two functions to develop a concurrent application. The traditional MapReduce framework initially faced challenges in supporting multiple datasets within a single task, but later advancements have addressed this issue [27, 28]. In order to improve programming efficiency, some frameworks with advanced language features have been proposed, providing efficient results, for example, Google's Sawzall [29], Yahoo's Pig Latin [30], Facebook's Hive [31] and Microsoft's SCOPE [32].

4.2.3.2. Dryad. A general-purpose distributed execution engine called Dryad [33] processes multiple applications with finely detailed data simultaneously.

The Dryad engine has an operational structure described by a directed acyclic graph, where programs are represented by vertices and data channels are represented by edges. Once a graph of operation cycle is in place, all resources involved in the logical operation graph are automatically mapped to physical resources. A central program called the job manager coordinates the operational structure of Dryad, which can be used in clusters or workstations through a network. The job manager comprises of two main parts: 1) application codes used to build a job communication graph, and 2) program library codes utilized to arrange available resources. Data transmission occurs between vertexes. Therefore, without obstructing any data transmission, the job manager is only responsible for decision-making. In Dryad, one can select any directed acyclic graph to describe application communication modes and data transmission mechanisms. The advanced language of Dryad, DryadLINQ [34], is utilized to combine an execution environment similar to SQL, as discussed previously.

4.2.3.3. All-pairs. The all-pairs [35] system is designed primarily for biometrics, bioinformatics, and data mining applications. All-pairs operates with three-components (set A, set B, and function F), where function F compares the two sets A and B. The comparison results produce a Cartesian product matrix M. The all-pairs system uses a four-phase system modelling, distribution of input data, batch job management and result collection. In the system modelling phase, an approximate model is built to measure the CPU resources needed and to perform job partition. In the second phase, a spanning tree for data transmission is constructed. In the third step, all-pairs engine then creates batch processing submission for the partitioned jobs. In the last step, once the job processing is complete, the result engine accumulates all the results and formulate a proper structured output.

4.2.3.4. Pregel. The Google Pregel system [36] facilitates graph analysis, such social network graphs, etc. Computational tasks in Pregel are represented through directed graphs. After defining all vertices and edges once the graph is in place, the program performs a series of iterative calculations termed as supersteps, globally synchronized points until the algorithm runs successfully and the output is delivered. During each superstep, vertex computations are performed adjacently and each vertex executes the same user-defined function to actualize the algorithm logic. All vertices have the power to alter the topological structure of the entire graph, change messages delivered from the previous superstep to other vertexes, change the status of their output edges, and more. Each edge corresponds uniquely to one vertex, ensuring all interactions are properly managed. The ability to remove individual vertex functions by suspension adds flexibility. When all vertexes are inactive and there is no message to transmit, the program

execution completes. The Pregel program output is defined as a set of all vertexes output values. Isomorphic directed graphs serve as both input and output of the Pregel program.

4.3. Big data analytics fields

Big data analytics involves examining large and varied datasets in order to gain insights out of that data. In this era of big data, the main focus is on discovering previously unknown facts. Finding insights from big data is so important that it has brought significant changes to many businesses, especially those handles mass customer bases. By using analytics on big data, businesses can understand the current state as well as customer behavior. Given the variety component of big data, which contains both structured and unstructured data, big data analytics encompasses the analysis of both types.

4.3.1. Structured data analysis. Structured data refers to the data that can be stored in fixed or predefined fields or records in a file. Every business organization, academic institution, and scientific research group generates massive amounts of structured data, which can be stored and analyzed by traditional database systems such as RDBMS, data warehouse, online analytical processing (OLAP) and business intelligence (BI) tools [78]. Data analysis, also known as data mining, has been an important field of study for over 40 years, and almost all of the business organizations use the tools of statistical analysis for better decision making and formulating future policies. Data analysis is still a very active field of research due to new applications demanding better mathematical or statistical models. Various business applications, such as credit risk management, fraud detection and customer loyalty analysis, require new models and powerful algorithms for their analysis.

4.3.2. Text mining. A vast amount of new data is generated every day from various economic, academic or social activities, and majority of this content is in the form of text such as emails, business documents, RSS feeds, data from social network sites, etc. Therefore, analyzing this textual data is deemed to hold more potential than analyzing structured data. Generally, text mining refers to extracting useful information and knowledge from unstructured data, and the field is very active and popular among research communities. There is too much literature available to scholars and researchers, with approximately 11 550 journals publishing around 1.5 million articles per year [79]. Text mining is interdisciplinary and involves various field such as information retrieval, machine learning, statistics and natural language processing (NLP). Recently, many

industries have begun leveraging text analysis to enhance their growth and to acquire competitive advantages. Some of the popular companies are adopting text analysis to support and improve their core and business activities. These companies include Netflix, Bank of England, IBM Watson's question-answering system, etc. [84]. With the increase of social networking sites such as Facebook, Twitter and the now-defunct Google+, text analysis expanded into new domains, widely used by several organizations and companies. This practice is known as social network analysis. The social network analysis covers many tasks such as sentiments analysis, classification, clustering, summarization and entity identification, among others. These domains are becoming increasingly popular and are used to analyze customer behavior, assess the popularity of newly launched products and predict future purchasing patterns of potential customers. Table 3 shows recent research work done in the field of text analysis and social network analysis.

4.3.3. Multimedia data analysis. Typically, multimedia data is categorized into two parts: one is structured and the other is semi-structured data. The multimedia data is stored in some specialized multimedia databases and mining it reveals some very interesting facts. Multimedia research is a popular research domain that helps in finding interesting facts and knowledge from multimedia datasets such as image, audio video, text, or combinations of all or some (Table 4).

4.3.4. Image mining. Image data represent a major keystone in various research domains such as medicine, forensics, criminology, robotics, meteorology and education. Extracting useful information from image database is very useful and can lead to some significant results. Unlike text data, images are stored and analyzed in a different way due to their different nature. Image mining refers to extracting useful information from image databases, such as implicit knowledge, image-data relationship, or some other patterns [80]. The field of image mining is interdisciplinary as it combines knowledge from many domains like machine learning, statistics, artificial intelligence, image processing, etc. Finding information from images represents a special data processing entity as images have visual characteristics and various features that can be represented numerically. Searching, classifying and analyzing images from image databases are basic requirements for image processing. Although tools for this task are available in data mining, they are not sufficient to provide efficient results in image processing. Various components that are involved in image mining include image analysis, object identification and recognition, image classification, image indexing and image retrieval.

Table 3. Recent works in the field of text mining.

Reference	Objective	Methodology	Strength	Weakness
Lavrenko *et al.* [93]	To propose a model that could identify news affecting markets trends	– Identified patterns in time series using piecewise linear fit model and automatic binning methods. – Used language model for language pattern associated with a trend.	Study behavior of market trends with news broadcasts	The results are not very appreciable
Thomas & Sycara [94]	To analyze the impact of textual information on financial market	– Proposed two models of prediction for financial market, one using maximum entropy algorithm for text classification and other using genetic algorithm to learn rule from numerical data. – Proposed a combined model for predicting future market trends.	Combined model predicts future market trends better than individual models	The proposed method lacks in handling noise in text data
Back *et al.* [95]	Compared results of quantitative and qualitative data from annual reports of the companies using clustering and neural networks	– Proposed a model using clustering and self-organizing maps (SOM). – Worked on numeric as well as textual data collected from organizations' annual reports.	The results of the model show the combined approach of clustering and SOM yields better results	The proposed method lacks in handling bots as they may lead to biased results
Fung *et al.* [96]	Analyzed and predicted stock prices based on text data and multiple time series	– Proposed a model that predicts stock market based on textual information gathered from news articles and the time series. – SVM classifier is used for the analysis.	The results show the model works well with the multiple time series data	The dataset is not exhaustive to generalize better results
Koppel & Shtrimberg [97]	Sentiment analysis is performed on news articles to perform stock prediction using SVM classifier	– Proposed a model for stock prediction using sentiment analysis. – SVM classifier is used.	The accuracy of 70% was reported	The accuracy of the model is quite low and there is scope of improvements

[Table 3. Cont.].

Reference	Objective	Methodology	Strength	Weakness
Dey *et al.* [98]	Studied the impact of financial news on stock market and proposes a framework for prediction	– Proposed a framework for stock market prediction and factors that affect the market. – LDA model is used for the identification of financial news.	The results show an accuracy of around 73%	Accuracy of the model is quite low and there is room for improvements
Wang *et al.* [99]	An ontology-based framework is proposed for investigative study to find the relationship between news and financial instruments	– Ontology based framework is proposed, which uses naïve Bayes algorithm for news sentiment analysis.	Combined approach of ontology and NB algorithm improves results	The proposed method failed to handle the bots
Nassirtoussi *et al.* [100]	A multi-layered architecture to predict financial market with the help of news headlines is proposed	– Multi-layer dimensionality reduction techniques are used. – Both semantic and sentiment analysis are used at different layers to predict the correct results.	The results were very good with an accuracy of 83.3% on real data	The in-depth analysis of the complete news has not been done so the influencing factors cannot be identified
Schafer *et al.* [101]	A study related to the effect of recommender system on e-commerce is conducted and the taxonomy for the recommender system is also proposed	– Studied the impact of recommender system on five popular e-commerce sites. – Identified five popular recommender system applications.	A theoretical study is conducted to determine the impact of recommender system and its applications	No specific results shown
Pang *et al.* [102]	A study on document classification using sentiment analysis is done using movie reviews	– Studied the three main sentiment classification algorithms: naïve Bayes, maximum entropy and support vector machine.	The results show that the algorithms perform better than the human-generated values.	The study lacks in showing the behavior of many other established algorithms and their comparison

[Table 3. Cont.].

Reference	Objective	Methodology	Strength	Weakness
Hu & Liu [103]	A technique for text summarization using customer reviews is I proposed	– Proposed a text summarization method based only upon the product features for which reviews has been given by customers.	The results show the summarized results are effective for product opinion mining	The method failed to work with imbalanced dataset
Popescu & Etzioni [104]	An opinion mining model known as OPINE is proposed for mining customer reviews on Amazon	– Proposed a new tool OPINE for finding sentiments of customers based on their reviews. – Performed experiment on Amazon reviews.	The results report high precision and recall values	The authors used a single dataset and the data size is quite small so it may fail to handle large data size
Bifet & Frank [105]	A Twitter sentiment analysis model is proposed and a comparative study of available models is performed	– Proposed a new sentiment analysis model for Twitter data streams. – Used naive Bayes stochastic gradient descent (SGD) and Hoeffding tree model to investigate the accuracy of the models.	The results shows that SGD model outperforms other available models with an accuracy of 82.8%	Comparison scope limited to specific algorithms; broader comparison needed
Dey *et al.* [106]	A framework for opinion mining is proposed using NLP and ontology modeling	– Proposed a framework that extracts useful information from noisy text and exact opinion from the freely available customer review text.	The model shows good accuracy	The dataset used is not exhaustive

Table 4. Various fields of big data mining.

Reference	Analysis type	Objective	Methodology
Devasena & Hemalatha [132]	Image mining	Mining a multimedia database for accurate image retrieval	LIM-based image matching approach is used with neural network
Rajendran & Madheswaran [133]	Image mining	A new technique of image classification is proposed for brain tumor classification using image mining	Pruned association rule mining using MARI algorithm is used
Krizhevsky *et al.* [134]	Image mining	An improved image classification network is developed using neural network on ILSVRC dataset	A convolution neural network is used for better results
Francois *et al.* [135]	Video mining	An ontology based framework is proposed for video event detection	Provide good results when used with domain ontologies and tools
Gargi *et al.* [136]	Video mining	A multi-stage scalable algorithm is proposed to discover video content from YouTube	A mix of pre-processing, graph clustering and post-processing algorithms is used, producing effective results
Zhang *et al.* [137]	Video mining	An approach of video classification is used to improve video taxonomy classification system	A graph clustering algorithm is used to improve classification accuracy

4.3.5. Audio mining. Music information retrieval is an emerging and active field of research, attracting growing attention from both academia and industry. Audio mining allows users to search or retrieve musical or audio content based not only on text but also by singing/playing or by means of sound experts, or by applying a query consisting of a combination of text and audio [81]. Audio classification and summarization are some of the fundamental research problems in this area and involve many grey areas such as genre classification, mood classification, and instrument and artist recognition, etc. Audio summarization and classification techniques use extracting words or phrases from metadata, and such methods are simple and are being used by many business applications.

4.3.6. Video mining. In recent years, the volume of video content has increased so much, with billions of videos being watched and uploaded daily on the web. For example, around 300 hours of videos are uploaded to YouTube every

minute and almost 5 billion videos are watched per day [82]. Therefore, content-based video analysis and retrieval are gaining international research interest. Since the volume of video data is so large, advanced and more robust techniques need to be developed for organizing, indexing, filtering, retrieving and mining such a content effectively.

5. Tools and techniques for big data

Many businesses are now able to handle big data effectively thanks to the most recent improvements in big data methodologies and technologies. With the emergence of a number of big data mining tools, it is now feasible to leverage the value of big data. Many organizations that handle big data use these tools for the leverage of big data and to expand their customer base. Many software companies have launched their data mining tools in the market, which can be used directly by other organizations to perform analytics on their data. Although the tools available in the market are not all free but still a number of open source tools and software are available, which helps in performing data mining tasks. Some of them are discussed below.

RapidMiner [85] is a software platform used by data scientists for performing data mining tasks. It provides an integrated environment for all the data mining tasks that include data preparation, machine learning models, deep learning, text analytics, etc. It is used by researchers for education and research and also by the businesses and organizations for commercial purposes. It is available under both open-source as well as with commercial pricing. RapidMiner uses a client-server architecture, and the server offers to serve on-premise, in public clouds or in private cloud infrastructure.

Weka [86] is a machine learning software suit written in Java. It was developed at the University of Waikato, New Zealand under the free GNU General Public License. Weka is popular among data science research community, as it provides visualization tools and algorithms for carrying out researches in data analysis and predictive modeling. Weka is developed in Java so it highly portable and it can run on nearly any computing platform.

Orange [87] is an open source data mining and machine learning toolkit developed at the University of Ljubljana. Orange is written in C++ with the wrappers of Python and Cython, and it provides interactive data visualization for their users with the help of a visual programming front-end. It is mainly used for machine learning algorithms and predictive modeling. Orange provides various widgets and add-ons for various algorithms and data mining tasks.

DataMelt (DMelt) [88] is a data analysis and data visualization framework used by data scientists, engineers and students. It is written in Java and it can

run on any platform supporting the Java virtual machine. DataMelt is an open source-framework that offers a unified user interface comparable to commercial software. It provides 2D and 3D visualization of data using histograms, charts and functions, and it is mainly used for numeric computations.

KEEL (Knowledge Extraction based on Evolutionary Learning) [89] is another open-source data mining tool. It is developed in Java and it provides a simple graphical user interface (GUI) for the users. KEEL contains a number of classical knowledge extraction algorithms, data pre-processing techniques, machine learning models and statistical methodologies supporting experimentation and comparison of different approaches. It is specifically designed for research and educational purposes.

SPMF (Sequential Pattern Mining Framework) [90] is an open-source data mining library developed in Java that contains more than 55 implemented data mining algorithms. It is a cross-platform and can run on any platform. It is designed specifically for pattern discovery. SPMF is specifically used for three types of mining problems: frequent itemset mining, association rule mining and sequential pattern mining. The source code of SPMF is available under the GNU General Public License.

Rattle [91] is an open-source data mining software package written in the R programming language. It provides a graphical user interface (GUI) for various data mining tasks and is available under the GNU General Public License. Rattle allows input from different file formats that include CSV, TXT, EXCEL, ARF, ODBC, RData files, etc. A large number of data mining and statistical algorithms is implemented in Rattle, which assist in various data analysis tasks and provide visualization of results in the form of charts and models. Rattle is used by various departments and organizations in Australia to perform data mining activities.

Apache Mahout is a project of the Apache Software Foundation providing a simple and extensible programming framework for scalable machine learning algorithms. It is written in Java and complimented with wrappers in Scala, Mahout provides core algorithms for collaborative filtering, classification and clustering and is particularly used for developing recommendation systems. Majority of its algorithms are implemented on the top of Hadoop framework.

Apart from the open-source tools and software, many commercially available tools are widely used by organizations and research communities. A list of popular big data mining tools is given below (Table 5).

5.1. Techniques

Several core techniques are used in data mining, mainly divided into two categories: supervised and unsupervised. In supervised learning, we have inputs

Table 5. List of tools available for big data analysis.

Tool name	Availability	Available link
Rapidminer	Both open source and paid	https://my.rapidminer.com/nexus/account/index.html#downloads
Weka	Open source	http://www.cs.waikato.ac.nz/ml/weka/downloading.html
Orange	Open source	https://orange.biolab.si/download/
DetaMelt	Open source	http://jwork.org/dmelt/index.php?id=install
KEEL	Open source	http://www.keel.es/
SPMF	Open source	http://www.philippe-fournier-viger.com/spmf/index.php?link=download.php
Rattle	Open source	http://rattle.togaware.com/rattle-download.html
Apache Mahout	Open source	https://mahout.apache.org/general/downloads.html
Dundas BI	Paid	http://www.dundas.com/dundas-bi
Domo	Paid	https://www.domo.com/pricing
IBM Cognos Analytics	Paid	https://www.ibm.com/analytics/us/en/technology/products/cognos-analytics/
Yellowfin	Paid	https://www.yellowfinbi.com/
BOARD	Paid	http://www.board.com/en
Pentaho	Paid	http://www.pentaho.com/
SAP Business Intelligence	Paid	https://www.sap.com/india/products/analytics/business-intelligence-bi.html

as well as outputs and we use algorithms that maps inputs to outputs. Supervised learning can be further grouped into two categories: classification and regression. In classification, the output is known and labelled, such as in case of email classification where the output classes may be 'spam' and 'not spam'. In case of regression, we predict the value of an unknown variable with the help of some known variables. We try to build a model (linear or non-linear) that fits our problem domain and maps our unknown values to the known values. In unsupervised learning, the labels of the output are not known in advance and it mainly includes clustering and anomaly detection. Clustering is the task of grouping similar types of objects into a single group or cluster. In clustering, the model learns the feature of the data and based upon similarity of features, the data is grouped into similar clusters. Anomaly detection, also known as outlier detection, is a technique of finding data items, events or observations that do not resemble with the expected output.

A list of available techniques (supervised and unsupervised) is given in the tables below (Tables 6 and 7, respectively).

Table 6. Supervised techniques for big data mining.

Reference	Technique	Uses
Wang [107]	Artificial neural network	Classification/regressions
Hecht-Nielsen [108]	Backpropagation	Classification
Freund *et al.* [109]	Boosting	Classification
Bernardo & Smith [110]	Bayesian statistics	Classification
Kolodner [111]	Case-based reasoning	Classification/regression
Suthaharan [112]	Decision tree learning	Classification
Muggleton *et al.* [113]	Inductive logic programming	Classification/regression
Choi [114]	Gaussian process regression	Regression
Murphy [115]	Naive Bayes classifier	Classification
McCallum *et al.* [116]	Maximum entropy classifier	Classification
Hearst *et al.* [117]	Support vector machines	Classification
Peterson [121]	k-nearest neighbor	Classification/regression

Table 7. Unsupervised techniques for big data mining.

Reference	Technique	Uses
Hartigan & Wong [118]	K-means	Clustering
Abu-Jamous *et al.* [119]	Mixture models	Clustering
Johnson [120]	Hierarchical clustering	Clustering
Ye & Chiang [122]	Apriori algorithm	Association rule mining
Schmidt-Thieme [123]	Eclat algorithm	Association rule mining
Borgelt [124]	FP-growth algorithm	Association rule mining

6. Challenges of big data and cloud computing

6.1. Data staging and pre-processing

Data staging and pre-processing is one of the important research issues when dealing with big data in a cloud computing environment. Due to the heterogeneity of data, the information gathered from various sources of big data includes both structured and unstructured components. For instance, data generated from several mobile cloud-based applications, microblogs, social networking sites, etc., generate data that contains text, images and videos. Pre-processing this data before analyzing it poses a difficult challenge for practitioners. Cleaning and transforming such heterogeneous data before loading it into data warehouses or other storage systems is a challenging task. Several efforts have been exerted to simplify this task of cleaning and transformation, such as Hadoop and MapReduce

to handle the unstructured data formats. However, we still need better solutions that can understand the context of unstructured data so that meaningful information can be retrieved for big data analysis.

6.2. Big data storage systems

In order to store big data, several solutions have been proposed, and some of them have been applied in the cloud computing environment. However, several issues still persist to have a simplified and efficient big data storage system that can also leverage the benefits of the cloud. One of the biggest hurdles in developing such a solution is the volume and velocity of big data. Due to the massive volume and the existing capabilities of cloud computing technologies to provide the necessary capacity and performance to address such volume, a number of opportunities exist for researchers and practitioners to develop such a solution. Nonetheless, this remains a significant challenge that needs to be addressed.

6.3. Big data analysis

Analyzing big data is highly important for any organization, and the selection of right model for analyzing big data is highly crucial as it varies from application to application. As the data volume increases very frequently, we need highly scalable algorithms to analyze big data that can easily handle massive data and deliver timely results. Existing algorithms fall short in this regard, so we need to develop more efficient. Also, big data contains unstructured data and streams arriving at high speeds from several sources, requiring integration with historical data to extract meaningful information. Available tools and techniques are not capable enough to handle it. Although researchers continue to probe algorithms and techniques that can easily manage and analyze big data in the cloud, the research is in its early stages, with ongoing efforts to develop integrated, efficient and effective solutions.

6.4. Security issues

With the advent of cloud computing, the modern ICT (Information and Communication Technology) industry has transformed significantly. However, cloud computing introduced several unresolved security threats and issues, such as privacy, availability, integrity and confidentiality. These security issues are amplified by the volume, velocity and variety of big data. So, security is a major concern for both cloud service providers and users when data is outsourced. Data cloud must be accessed and verified at regular intervals to ensure its security and privacy. Cloud vendors and service providers must ensure that the service level agreement is fully met. Some controversies have recently shown that the security

and privacy of the data were compromised. One of the reasons for such cases is due to the lack of proper security measures and the available techniques which are not capable to protect the data in cloud. Also we lack policies that cover all privacy concerns of users. A proper benchmark is necessary to ensure the privacy of data in the cloud. Strong cryptographic algorithms are needed to encrypt sensitive data in the cloud, ensuring secure key management and data access. Furthermore, traditional approaches like hashing for maintaining data integrity face challenges due to the sheer volume of data.

A number of other challenges also exist, given below, which need to be addressed in future researches so that big data and cloud computing can be combined and leveraged (Table 8).

Table 8. Other challenges of big data in cloud.

Reference	Issue	Description
Schroek *at al.* [138]	Availability	It refers to the on-demand accessibility of data. In case of big data stored on the cloud, one of the issues is the data availability. As the number of users increases in cloud, the service provider must address this issue. With the increase in streaming data generated by the organizations, the business model will demand fast access to data in more real-time.
Sravan Kumar & Saxena [139]	Data integrity	It refers to the correctness of data stored in the cloud. It is one of significant issues to maintain data integrity in the cloud because users may not be able to access data physically. Therefore, more robust and proper mechanisms are required so that only authorized users can modify data, and data integrity is preserved.
Kocarev & Jakimoski [140]	Data heterogeneity	Big data contains data in various formats (structured and unstructured) because data is collected from many sources. When managing big data in the cloud, it is a challenging task to handle this data heterogeneity.
Weber *et al.* [141]	Data quality	Since big data is collected from many heterogeneous sources, the quality of data can be compromised. To gain better insights from big data analysis, the data should be of good quality and it is a challenging task to obtain good quality data from different sources.
Akerkar [142]	Data transformation	Due to the variety of data present in large volumes, data transformation to a suitable form that can be easily used for analysis is a challenging task in case of big data.

7. Conclusion

Currently, the size of data is enormous and increases day by day due to the proliferation of mobile devices, sensors and other devices connected to the internet. Also, the velocity and variety of data are increasing, altering the traditional paradigm of data management and giving rise to a new paradigm known as big data. This new big data paradigm provides several opportunities for businesses across all industries to gain insights on real-time basis. With the advent of cloud computing used to store, process and analyze data, the landscape of the IT industry has changed into an on-demand service model industry. In this study, we also presented a review of big data and cloud computing and their co-existence in many applications. We presented many available definitions of big data and cloud computing and their uses in various fields. We also presented various cloud service models and their applications in various domains. We studied in detail the lifecycle of big data, starting from data generation sources, big data storage solutions and concluded with data analysis techniques that serve diverse purposes. We also studied various big data analysis schemes that have many applications in various applications. At the end, we presented several open research issues that need to be addressed and could provide directions for new researches in the fields of big data and cloud computing.

References

1. D. Laney, 3-D data management: Controlling data volume, velocity and variety, META Group Research Note 6, 2001, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
2. H. Chen, R.H.L. Chiang, V.C. Storey, Business intelligence and analytics: From big data to big impact, *MIS Quarterly*, **36**(4): 1165–1188, 2012, doi: 10.2307/41703503.
3. O. Kwon, N. Lee, B. Shin, Data quality management, data usage experience and acquisition intention of big data analytics, *International Journal of Information Management*, **34**(3): 387–394, 2014, doi: 10.1016/j.ijinfomgt.2014.02.002.
4. Gartner, IT Glossary, Big Data, n.d., http://www.gartner.com/it-glossary/big-data/.
5. D. Beaver, S. Kumar, H.C. Li, J. Sobel, P. Vajgel, Finding a needle in haystack: Facebook's photo storage, [in:] Proceedings of the Ninth USENIX Symposium on Operating Systems Design and Implementation (OSDI 10), Berkeley, CA, USA, pp. 1–8, USENIX Association, 2010, https://research.facebook.com/publications/finding-a-needle-in-haystack-facebooks-photo-storage/.
6. K. Cukier, Data, data everywhere: A special report on managing information, *The Economist*, February 25, 2010, http://www.economist.com/node/15557443.
7. Y. Demchenko, P. Grosso, C. de Laat, P. Membrey, Addressing big data issues in scientific data infrastructure, [in:] *2013 International Conference on Collaboration Technologies and Systems (CTS)*, San Diego, CA, USA, pp. 48–55, 2013, doi: 10.1109/CTS.2013.6567203.

8. A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, **35**(2): 137–144, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

9. J. Manyika *et al.*, Big data: The next frontier for innovation, competition, and productivity, Report, McKinsey Global Institute, 2011, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

10. C.L.P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, *Information Sciences*, **275**: 314–347, 2014, doi: 10.1016/j.ins.2014.01.015.

11. L. Candela, D. Castelli, P. Pagano, Managing big data through hybrid data infrastructures, *ERCIM News*, **89**: 37–38, 2012.

12. J. Gantz, D. Reinsel, Extracting value from chaos, *IDC' Digital Universe Study, IDC iView*, pp. 1–12, 2011.

13. Fact sheet: Big data across the federal government, The White House, March 29, 2012. https://obamawhitehouse.archives.gov/the-press-office/2015/12/04/fact-sheet-big-data-across-federal-government.

14. V. Mayer-Schönberger, K. Cukier, *Big data: A Revolution that Will Transform how We Live, Work, and Think*, An Eamon Dolan Book/Houghton Mifflin Harcourt, Boston, New York, 2013.

15. M. Chen, S. Mao, Y. Liu, Big data: A survey, *Mobile Networks and Applications*, **19**(2): 171–209, 2014, doi: 10.1007/s11036-013-0489-0.

16. O'Reilly Radar Team, *Big Data Now: Current Perspectives from O'Reilly Radar*, O'Reilly Media, 2011.

17. M. Grobelnik, Big data tutorial, 2012, http://videolectures.net/eswc2012grobelnikbigdata/ (accessed May 12, 2017).

18. A. Labrinidis, H.V. Jagadish, Challenges and opportunities with big data, *Proceedings of the VLDB Endowment*, **5**(12): 2032–2033, 2012, doi: 10.14778/2367502.2367572.

19. PoweredBy – Applications and organizations using HADOOP2, Apache Software Foundation, 2013, http://wiki.apache.org/hadoop/PoweredBy.

20. T. Gunarathne, T.-L. Wu, J.Y. Choi, S.-H. Bae, J. Qiu, Cloud computing paradigms for pleasingly parallel biomedical applications, *Concurrency and Computation: Practice and Experience*, **23**(17): 2338–2354, 2011, doi: 10.1002/cpe.1780.

21. J. Gantz, D. Reinsel, The digital universe decade – Are you ready?, *IDC Analyze the Future*, pp. 1–16, 2010.

22. *How Big Data Analysis helped increase Walmart's Sales turnover?*, ProjectPro, https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109 (accessed May 12, 2017).

23. R. Cattell, Scalable SQL and NoSQL data stores, *ACM SIGMOD Record*, **39**(4): 12–27, 2011, doi: 10.1145/1978915.1978919.

24. E. Ma, *Colossus: Successor to the Google File System (GFS)*, SysTutorials, https://www.systutorials.com/colossus-successor-to-google-file-system-gfs/ (accessed May 12, 2017).

25. R. Chaiken *et al.*, SCOPE: Easy and efficient parallel processing of massive data sets, *Proceedings of the VLDB Endowment*, **1**(2): 1265–1276, 2008, doi: 10.14778/1454159.1454166.

26. J. Dean, S. Ghemawat, MapReduce: Simplified data processing on large clusters, *Communications of the ACM*, **51**(1): 107–113, 2008, doi: 10.1145/1327452.1327492.

27. S. Blanas, J.M. Patel, V. Ercegovac, J. Rao, E.J. Shekita, Y. Tian, A comparison of join algorithms for log processing in MapReduce, [in:] *SIGMOD'10: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, pp. 975–986, ACM, 2010, doi: 10.1145/1807167.180727.

28. H.-C. Yang, D.S. Parker, Traverse: Simplified indexing on large Map-Reduce-Merge clusters, [in:] *Database Systems for Advanced Applications*, X. Zhou, H. Yokota, K. Deng, Q. Liu [Eds.], Springer, pp. 308–322, 2009.

29. R. Pike, S. Dorward, R. Griesemer, S. Quinlan, Interpreting the data: Parallel analysis with Sawzall, *Scientific Programming*, **13**(4): 277–298, 2005, doi: 10.1155/2005/962135.

30. A.F. Gates *et al.*, Building a high-level dataflow system on top of Map-Reduce: The Pig experience, *Proceedings of VLDB Endowment*, **2**(2): 1414–1425, 2009, doi: 10.14778/1687553.1687568.

31. A. Thusoo *et al.*, Hive: A warehousing solution over a Map-Reduce framework, *Proceedings of the VLDB Endowment*, **2**(2): 1626–1629, 2009, doi: 10.14778/1687553.1687609.

32. M.-C. Wu, J. Zhou, N. Bruno, Y. Zhang, J. Fowler, Scope playback: Self-validation in the cloud, [in:] *Proceedings of the Fifth International Workshop on Testing Database Systems (DBTest'12)*, Article 3, pp. 1–6, Association for Computing Machinery, New York, NY, USA, 2012, doi: 10.1145/2304510.2304514.

33. M. Isard, M. Budiu, Y. Yu, A. Birrell, D. Fetterly, Dryad: Distributed data-parallel programs from sequential building blocks, *ACM SIGOPS Operating Systems Review*, **41**(3): 59–72, 2007, doi: 10.1145/1272996.1273005.

34. Y. Yu *et al.*, DryadLINQ: A system for general-purpose distributed data-parallel computing using a high-level language, [in:] *8th USENIX Symposium on Operating Systems Design and Implementation*, San Diego, CA, USA, Vol. 8, pp. 1–14, 2008.

35. C. Moretti, J. Bulosan, D. Thain, P.J. Flynn, All-pairs: An abstraction for data-intensive cloud computing, [in:] *2008 IEEE International Symposium on Parallel and Distributed Processing*, Miami, FL, USA, 2008, pp. 1–11, doi: 10.1109/IPDPS.2008.4536311.

36. G. Malewicz *et al.*, Pregel: A system for large-scale graph processing, [in:] *SIGMOD'10: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, pp. 135–146, ACM, 2010, doi: 10.1145/1807167.180718.

37. C.-W. Lu, C.-M. Hsieh, C.-H. Chang, C.-T. Yang, An improvement to data service in cloud computing with content sensitive transaction analysis and adaptation, [in:] *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*, Japan, pp. 463–468, 2013, doi: 10.1109/COMPSACW.2013.72.

38. M. Armbrust *et al.*, A view of cloud computing, *Communications of the ACM*, **53**(4): 50–58, 2010, doi: 10.1145/1721654.1721672.

39. H. Liu, Big data drives cloud adoption in enterprise, *IEEE Internet Computing*, **17**(4): 68–71, 2013, doi: 10.1109/MIC.2013.63.

40. S. Pandey, S. Nepal, Cloud computing and scientific applications – Big data, scalable analytics, and beyond, *Future Generation Computer Systems*, **29**(7): 1774–1776, 2013, doi: 10.1016/j.future.2013.04.026.

41. D. Warneke, O. Kao, Nephele: Efficient parallel data processing in the cloud, [in:] *MTAGS'09: Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers*, ACM, Article no. 8, pp. 1–10, 2009, doi: 10.1145/1646468.1646476.

42. P. Mell, T. Grance, *The NIST Definition of Cloud Computing*, Technical Report, Special Publication 80, National Institute of Standards & Technology, Gaithersburg, MD, USA, 2011.

43. G. Aceto, A. Botta, W. de Donato, A. Pescapè, Cloud monitoring: A survey, *Computer Networks*, **57**(9): 2093–2115, 2013, doi: 10.1016/j.comnet.2013.04.0.

44. T. Gunarathne, B. Zhang, T.-L. Wu, J. Qiu, Scalable parallel computing on clouds using Twister4Azure iterative MapReduce, *Future Generation Computer Systems*, **29**(4): 1035–1048, 2013, doi: 10.1016/j.future.2012.05.027.

45. A. O'Driscoll, J. Daugelaite, R.D. Sleator, 'Big data', Hadoop and cloud computing in genomics, *Journal of Biomedical Informatics*, **46**(5): 774–781, 2013, doi: 10.1016/j.jbi.2013.07.001.

46. M.D. Assunção, R.N. Calheiros, S. Bianchi, M.A.S. Netto, R. Buyya, Big Data computing and clouds: Trends and future directions, *Journal of Parallel and Distributed Computing*, 79–80: 3–15, 2015, doi: 10.1016/j.jpdc.2014.08.003.

47. P.S. Yu, On mining big data, [in:] J. Wang, H. Xiong, Y. Ishikawa, J. Xu, J. Zhou [Eds.], *Web-Age Information Management*, Lecture Notes in Computer Science, Vol. 7923, Springer-Verlag, Berlin, Heidelberg, 2013, p. XIV.

48. X. Sun *et al.*, Towards delivering analytical solutions in cloud: Business models and technical challenges, [in:] *2011 IEEE 8th International Conference on e-Business Engineering*, Beijing, China, pp. 347–351, 2011, doi: 10.1109/ICEBE.2011.81.

49. '*Big Data' has Big Potential to Improve Americans' Lives, Increase Economic Opportunities*, Press Releases, Committee on Science, Space and Technology, April 24, 2013, https://science.house.gov/2013/4/big-data-has-big-potential-improve-americans-lives-increase-economic.

50. Prime Minister joins Sir Ka-shing Li for launch of £90m initiative in big data and drug discovery at Oxford, University of Oxford, May 3, 2013, http://www.cs.ox.ac.uk/news/639-full.html.

51. J. Manzoni, Big data in government: the challenges and opportunities, Speech delivered on February 17, 2017, https://www.gov.uk/government/speeches/big-data-in-government-the-challenges-and-opportunities.

52. Government-backed Russian Fund Launches Big Data Investment Program, *RusSoft*, http://russoft.org/docs/?doc=3391 (accessed May 12, 2017).

53. bigdata@csail, http://bigdata.csail.mit.edu/ (accessed May 12, 2017).

54. The Intel science and technology center for big data, Information Science and Technology Consultants (ISTC), http://istc-bigdata.org.

55. D. Borthakur *et al.*, Apache Hadoop goes realtime at Facebook, [in:] *SIGMOD'11: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, New York, USA, pp. 1071–1080, ACM, 2011, doi: 10.1145/1989323.1989438.

56. M. Armbrust *et al.*, *Above the Clouds: A Berkeley View of Cloud Computing*, Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, USA, 2009, http://www2.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf.

57. Google. Google Trends for Big Data, 2013.

58. R. Swanstrom, NIST defines Big Data and Data Science, Data Science 101, Learn Data Science, http://101.datascience.community/2015/04/23/nist-defines-big-data-and-data-science/ (accessed May 12, 2017).

59. B. Golden, *Virtualization for Dummies*, Wiley, Indianapolis, Indian, USA, 2009.

60. Zephoria Inc., The Top 20 Valuable Facebook Statistics – Updated May 2017, https://zephoria.com/top-15-valuable-facebook-statistics/ (accessed May 12, 2017).

61. Y. Yang, L. Zhang, Y. Zhen, R. Ji, Learning for visual semantic understanding in big data, *Neurocomputing*, **169**: 1–4, 2015, doi: 10.1016/j.neucom.2015.05.023.

62. S. Maitrey, C.K. Jha, MapReduce: Simplified data analysis of big data, *Procedia Computer Science*, **57**: 563–571, 2015, doi: 10.1016/j.procs.2015.07.392.

63. A. Vinay, V.S. Shekhar, J. Rituparna, T. Aggrawal, K.N.B. Murthy, S. Natarajan, Cloud based big data analytics framework for face recognition in social networks using machine learning, *Procedia Computer Science*, **50**: 623–630, 2015, doi: 10.1016/j.procs.2015.04.095.

64. K.P.N. Jayasena, L. Li, Q. Xie, Multi-modal multimedia big data analyzing architecture and resource allocation on cloud platform, *Neurocomputing*, **253**: 135–143, 2017, doi: 10.1016/j.neucom.2016.11.077.

65. C. Smowton *et al.*, A cost-effective approach to improving performance of big genomic data analyses in clouds, *Future Generation Computer Systems*, **67**: 368–381, 2017, doi: 10.1016/j.future.2015.11.011.

66. D. Linthicum, Three types of IoT data sources, RTInsights.com, March 29, 2016, https://www.rtinsights.com/three-types-of-iot-data-sources/ (accessed May 12, 2017).

67. M. Wilhelm *et al.*, Mass-spectrometry-based draft of the human proteome, *Nature*, **509**: 582–587, 2014, doi: 10.1038/nature13319.

68. E. Olshannikova, T. Olsson, J. Huhtamäki, H. Kärkkäinen, Conceptualizing big social data, *Journal of Big Data*, **4**(3): 1–19, 2017.

69. A.E. Marwick, *Status Update: Celebrity, Publicity, and Branding in the Social Media Age*, Yale University Press, 2013.

70. F. Campos Freire, N. Alonso Ramos, Online digital social tools for professional self-promotion. A state of the art review, *Revista Latina de Comunicación Social*, **70**: 288–299, 2015, doi: 10.4185/RLCS-2015-1047en.

71. C. Shih, *The Facebook Era: Tapping Online Social Networks to Build Better Products, Reach New Audiences, and Sell More Stuff*, Prentice Hall, New York, 2009.

72. A.T. Stephen, O. Toubia, Deriving value from social commerce networks, *Journal of Marketing Research*, **47**(2): 215–228, 2010, doi: 10.2139/ssrn.1150995.

73. M.T. Musacchio, R. Panizzon, X. Zhang, V. Zorzi, A linguistically-driven methodology for detecting impending disasters and unfolding emergencies from social media messages, [in:] *Proceedings of the LREC 2016 Workshop "EMOT: Emotions, Metaphors, Ontology*

and Terminology during Disasters, K. Ahmad, S. Kelly, X. Zhang [Eds.], Portorož, Slovenia, p. 26–33, 2016.

74. C. Aradau, T. Blanke, Politics of prediction: Security and the time/space of governmentality in the age of big data, *European Journal of Social Theory*, **20**(3): 373–391, 2017, doi: 10.1177/1368431016667623.
75. A.M.M. Saldana-Perez, M. Moreno-Ibarra, Traffic analysis based on short texts from social media, *International Journal of Knowledge Society Research (IJKSR)*, **7**(1): 63–79, 2016, doi: 10.4018/IJKSR.2016010105.
76. E. Qualman, *Socialnomics: How Social Media Transforms the Way We Live and Do Business*, John Wiley & Sons, New Jersey, 2010.
77. H. Kennedy, Commercial mediations of social media data, [in:] *Post, Mine, Repeat*, pp. 99–127, Palgrave Macmillan, London, 2016, doi: 10.1057/978-1-137-35398-6_5.
78. D. Agrawal *et al.*, *Challenges and Opportunities with Big Data*, A white paper prepared for the Computing Community Consortium Committee of the Computing Research Association, 2012, http://cra.org/ccc/resources/ccc-led-whitepapers.
79. M. Ware, M. Mabe, *The STM Report: An Overview of Scientific and Scholarly Journal Publishing*, International Association of Scientific, Technical and Medical Publishers, The Hauge, The Netherlands, 2009.
80. M.C. Burl, C. Fowlkes, J. Roden, Mining for image content, [in:] *Systemics, Cybernetics, and Informatics/Information Systems: Analysis and Synthesis, Session on Intelligent Data Mining and Knowledge Discovery*, 1999.
81. N. Kennedy, *Facebook's photo storage rewrite*, https://www.niallkennedy.com/blog/2009/04/facebook-haystack.html (accessed May 12, 2017).
82. FortuneLords, YouTube Statistics – 2017, https://fortunelords.com/youtube-statistics/ (accessed May 12, 2017).
83. D. Saravanan, S. Srinivasan, Data mining framework for video data, [in:] *Recent Advances in Space Technology Services and Climate Change 2010 (RSTS & CC-2010)*, Chennai, India, pp. 167–170, 2010, doi: 10.1109/RSTSCC.2010.5712827.
84. A. Ittoo, L.M. Nguyen, A. van den Bosch, Text analytics in industry: Challenges, desiderata and trends, *Computers in Industry*, **78**: 96–107, 2016, doi: 10.1016/j.compind.2015.12.001.
85. RapidMiner, https://rapidminer.com/ (accessed May 12, 2017).
86. Weka, http://www.cs.waikato.ac.nz/ml/weka/ (accessed May 12, 2017).
87. Orange Data Mining, https://orange.biolab.si/ (accessed May 12, 2017).
88. DataMelt, http://jwork.org/dmelt/ (accessed May 12, 2017).
89. KEEL, http://www.keel.es/ (accessed May 12, 2017).
90. P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, Ch.-W. Wu, V.S. Tseng, SPMF: A Java open-source pattern mining library, *Journal of Machine Learning Research*, **15**(1): 35699–3573, 2014.
91. G.J. Williams, Rattle: A data mining GUI for R, *The R Journal*, **1/2**: 45–55, 2009.
92. Apache Mahout, http://mahout.apache.org/ (accessed May 12, 2017).

93. V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, Mining of concurrent text and time series, [in:] *KDD-2000 Workshop on Text Mining*, Vol. 2000, pp. 37–44, University Park, PA, USA, 2000.

94. J.D. Thomas, K. Sycara, Integrating genetic algorithms and text learning for financial prediction, [in:] *Proceedings of GECCO'00 Workshop on Data Mining with Evolutionary Algorithms*, pp. 72–75, 2000.

95. B. Back, J. Toivonen, H. Vanharanta, A. Visa, Comparing numerical data and text information from annual reports using self-organizing maps, *International Journal of Accounting Information Systems*, **2**(4): 249–269, 2001, doi: 10.1016/S1467-0895(01)00018-5.

96. G.P.C. Fung, J.X. Yu, W. Lam, Stock prediction: Integrating text mining approach using real-time news, [in:] *2003 IEEE International Conference on Computational Intelligence for Financial Engineering*, Hong Kong, China, pp. 395–402, 2003, doi: 10.1109/CIFER.2003.1196287.

97. M. Koppel, I. Shtrimberg, Good news or bad news? Let the market decide, [in:] *Computing Attitude and Affect in Text: Theory and Applications*, J.G. Shanahan, Y. Qu, J. Wiebe [Eds.], pp. 297–301, Springer, Dordrecht, 2006, doi: 10.1007/1-4020-4102-0_22.

98. L. Dey, A. Mahajan, S.K.M. Haque, Document clustering for event identification and trend analysis in market news, [in:] *ICAPR '09: Proceedings of the 2009 Seventh International Conference on Advances in Pattern Recognition*, pp. 103–106, Kolkata, India, 2009, doi: 10.1109/ICAPR.2009.84.

99. S. Wang, K. Xu, L. Liu, B. Fang, S. Liao, H. Wang, An ontology based framework for mining dependence relationships between news and financial instruments, *Expert Systems with Applications*, **38**(10): 12044–12050, 2011, doi: 10.1016/j.eswa.2011.01.148.

100. A.K. Nassirtoussi, S. Aghabozorgi, T.Y. Wah, D.C.L. Ngo, Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment, *Expert Systems with Applications*, **42**(1): 306–324, 2015, doi: 10.1016/j.eswa.2014.08.004.

101. J.B. Schafer, J.A. Konstan, J. Riedl, E-commerce recommendation applications, *Data Mining and Knowledge Discovery*, **5**(1): 115–153, 2001, doi: 10.1023/A:1009804230409.

102. B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, [in:] *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 79–86, Association for Computational Linguistics, 2002, doi: 10.3115/1118693.1118704.

103. M. Hu, B. Liu, Mining and summarizing customer reviews, [in:] *KDD '04: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004, doi: 10.1145/1014052.1014073.

104. A.-M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, [in:] *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 339–346, Association for Computational Linguistics, 2005, https://aclanthology.org/H05-1043.

105. A. Bifet, E. Frank, Sentiment knowledge discovery in Twitter streaming data, [in:] B. Pfahringer, G. Holmes, A. Hoffmann [Eds.], *Discovery Science*, Lecture Notes in Computer Science, Vol. 6332, pp. 1–15, Springer, Berlin, Heidelberg, 2010, doi: 10.1007/978-3-642-16184-1_1.

106. L. Dey, S.M. Haque, N. Raj, Mining customer feedbacks for actionable intelligence, [in:] *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, ON, Canada, pp. 239–242, 2010, doi: 10.1109/WI-IAT.2010.196.

107. S.-C. Wang, Artificial neural network, [in:] *Interdisciplinary Computing in Java Programming*, The Springer International Series in Engineering and Computer Science, Vol. 743, pp. 81–100, Springer, Boston, MA, 2003, doi: 10.1007/978-1-4615-0377-4_5.

108. R. Hecht-Nielsen, Theory of the backpropagation neural network, *Neural Networks for Perception*, **1**(Supplement 1): 445–448, 1988, doi: 10.1016/0893-6080(88)90469-8.

109. Y. Freund, R. Iyer, R.E. Schapire, Y. Singer, An efficient boosting algorithm for combining preferences, *Journal of Machine Learning Research*, **4**(6): 933–969, 2003.

110. J.M. Bernardo, A.F.M. Smith, *Bayesian Theory*, John Wiley and Sons, 2001.

111. J. Kolodner, *Case-Based Reasoning*, Morgan Kaufmann, 1993.

112. S. Suthaharan, Decision tree learning, [in:] *Machine Learning Models and Algorithms for Big Data Classification*, Integrated Series in Information Systems, Vol. 36, p. 237–269, Springer, Boston, MA, 2016, doi: 10.1007/978-1-4899-7641-3_10.

113. S. Muggleton, R. Otero, A. Tamaddoni-Nezhad [Eds.], *Inductive Logic Programming*, Vol. 38, Academic Press, London, 1992.

114. S. Choi, *Gaussian Process Regression Analysis for Functional Data*, Taylor & Francis, 2011.

115. K.P. Murphy, *Naive Bayes Classifiers*, University of British Columbia, 2006.

116. A. McCallum, D. Freitag, F.C.N. Pereira, Maximum entropy Markov models for information extraction and segmentation, [in:] *17th International Conference on Machine Learning*, Vol. 17, pp. 591–598, 2000.

117. M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, *IEEE Intelligent Systems and their Applications*, **13**(4): 18–28, 1998, doi: 10.1109/5254.708428.

118. J.A. Hartigan, M.A. Wong, Algorithm AS 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1): 100–108, 1979, doi: 10.2307/2346830.

119. B. Abu-Jamous, R. Fa, A.K. Nandi, Mixture model clustering, [in:] *Integrative Cluster Analysis in Bioinformatics*, Ch. 15, pp. 197–226, 2015, doi: 10.1002/9781118906545.ch15.

120. S.C. Johnson, Hierarchical clustering schemes, *Psychometrika*, **32**(3): 241–254, 1967, doi: 10.1007/BF02289588.

121. L.E. Peterson, K-nearest neighbor, *Scholarpedia*, **4**(2): 1883, 2009, doi: 10.4249/scholarpedia.1883.

122. Y. Ye, C.-C. Chiang, A parallel apriori algorithm for frequent itemsets mining, [in:] *Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06)*, Seattle, WA, USA, pp. 87–94, 2006, doi: 10.1109/SERA.2006.6.

123. L. Schmidt-Thieme, Algorithmic features of Eclat, [in:] *FIMI'04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, Brighton, UK, 2004.

124. C. Borgelt, An implementation of the FP-growth algorithm, [in:] *OSDM '05: Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, pp. 1–5, ACM, 2005, doi: 10.1145/1133905.1133907.

125. Cloud Computing Services – Amazon Web Services (AWS), https://aws.amazon.com/ (accessed May 12, 2017).

126. GoGrid, https://www.datapipe.com/gogrid/ (accessed May 12, 2017).

127. Flexiscale, http://www.flexiscale.com/signup-on-stop/ (accessed May 12, 2017).

128. App Engine Application Platform | Google Cloud, https://cloud.google.com/appengine/ (accessed May 12, 2017).

129. Cloud Computing Services | Microsoft Azure, https://azure.microsoft.com/en-in/ (accessed May 12, 2017).

130. RightScale, http://www.rightscale.com/ (accessed May 12, 2017).

131. Eucalyptus, http://www.dxc.technology/cloud/offerings/140041/140149-eucalyptus_software_support_services (accessed May 12, 2017).

132. C.L. Devasena, M. Hemalatha, A hybrid image mining technique using LIM-based data mining algorithm, *International Journal of Computer Applications*, **25**(2): 11–15, 2011, doi: 10.5120/3007-4056.

133. P. Rajendran, M. Madheswaran, An improved image mining technique for brain tumour classification using efficient classifier, *arXiv*, 2010, arXiv: 10.48550/arXiv.1001.1988.

134. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Communications of the ACM*, **60**(6): 84–90, 2017, doi: 10.1145/3065386.

135. A.R.J. Francois, R. Nevatia, J. Hobbs, R.C. Bolles, J.R. Smith, VERL: An ontology framework for representing and annotating video events, *IEEE Multimedia*, **12**(4): 76–86, 2005, doi: 10.1109/MMUL.2005.87.

136. U. Gargi, W. Lu, V. Mirrokni, S. Yoon, Large-scale community detection on YouTube for topic discovery and exploration, [in:] *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5, No. 1, pp. 486–489, 2011, doi: 10.1609/icwsm.v5i1.14191.

137. J.R. Zhang, Y. Song, T. Leung, Improving video classification via YouTube video co-watch data, [in:] *SBNMA'11: Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access*, pp. 21–26, ACM, 2011, doi: 10.1145/2072627.2072635.

138. M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, P. Tufano, *Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data. Executive Report*, IBM Institute for Business Value, 2012, https://public.dhe.ibm.com/software/uk/data/pdf/The_real-world_use_of_big_data.pdf.

139. R. Sravan Kumar, A. Saxena, Data integrity proofs in cloud storage, [in:] *2011 Third International Conference on Communication Systems and Networks (COMSNETS 2011)*, Bangalore, India, pp. 1–4, 2011, doi: 10.1109/COMSNETS.2011.5716422.

140. L. Kocarev, G. Jakimoski, Logistic map as a block encryption algorithm, *Physics Letters A*, **289**(4–5): 199–206, 2001, doi: 10.1016/S0375-9601(01)00609-0.

141. K.L. Weber, G. Rincon, A.L. Van Eenennaam, B.L. Golden, J.F. Medrano, Differences in allele frequency distribution of bovine high-density genotyping platforms in holsteins and jerseys, [in:] *Proceedings, Western Section, American Society of Animal Science*, Vol. 63, pp. 70–74, 2012, https://www.asas.org/docs/western-section/wsasas_2012.pdf?sfvrsn=0#page=84.

142. R. Akerkar [Ed.], *Big Data Computing*, Chapman and Hall/CRC Press, New York, 2013.

143. Statista, Big data market size revenue forecast worldwide from 2011–2027, https://www.statista.com/statistics/254266/global-big-data-market-forecast/ (accessed May 12, 2017).

144. S. Kumar, K. Cengiz, S. Vimal, A. Suresh, Energy efficient resource migration based load balance mechanism for high traffic applications IoT, *Wireless Personal Communications*, **127**: 385–403, 2021, doi: 10.1007/s11277-021-08269-7.

145. S. Kumar, P. Ranjan, R. Ramaswami, M.R. Tripathy, Resource efficient clustering and next hop knowledge based routing in multiple heterogeneous wireless sensor networks, *International Journal of Grid and High Performance Computing*, **9**(2): 1–20, 2017, doi: 10.4018/IJGHPC.2017040101.

146. S. Kumar, P. Ranjan, R. Radhakrishnan, M.R. Tripathy, Energy efficient multichannel MAC protocol for high traffic applications in heterogeneous wireless sensor networks, *Recent Advances in Electrical and Electronic Engineering*, **10**(3): 223–232, 2017, doi: 10.2174/2352096510666170601090202.

Received September 29, 2022; revised version February 27, 2023;
accepted March 10, 2023; published online August 14, 2024.