Big data: definition, characteristics, life cycle, applications, and challenges

Hiba Basim Alwan1 and Ku Ruhana Ku-Mahamud2

1 Department of Computer Engineering, Al-Mansour University College, 10068 Al-Andalus Sq., Baghdad, Iraq
2 School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

hiba.basim@muc.edu.iq

Abstract. Any data set contains large volumes of information and complex data is called Big Data (BD). BD is unlike other traditional data sets, so it requires special processing to manage it. BD faces many challenges starting from data capture through to the final results. BD exists in many subject areas such as business, governments, sciences, healthcare and transport. Thus it touches peoples’ lives in many aspects. BD is considered as the most important topic and requires good understanding in order to be fully utilized. This paper presents the basic information of BD which includes its properties and applications. Descriptions and examples of BD and its categories are elaborated upon. The BD architectural establishment is presented followed by the conclusion of the importance of BD.

1. Introduction

One of the greatest trendy ideas these days is Big Data (BD). Everyone speaks about BD as can be seen in the media. Governments and businesses attempt to use and implement BD to their benefits [1]. The term BD was not known until in the middle of 2011. Like cloud computing the term has been implemented from product sellers to huge scale outsourcing and cloud service suppliers to powerfully encourage their offerings [2]. But what actually is BD?

Lisa [3] defines BD as a group of data from conventional and digital bases inside/outside the enterprise that characterizes a basis for continuing detection and analysis. Another definition of BD is found in [4] which defined BD as a quantity of data that is extremely relative and cannot be managed through the usual methods while [5] defined BD as not only one technology, but a group of old and modern technologies that assist businesses to obtain actionable perceptions. BD has large quantities of dissimilar data which allows processing in real time analysis and response. BD can be defined as a giant size of data [6]. Within BD, analyzing, visualizing or any other processing can be done. From the above definitions, it can be concluded that if data cannot be stored or processed by a common system’s capabilities or exceed a common system’s capabilities then these data are considered BD. The powerful services in this modern domain are continuously growing volumes of data and the advances in technologies are always able to mine the data for commercial purposes [4].

The unexpected increase of BD like a modern resource of knowledge has encouraged business decision-makers to generate decisions more quickly and to proactively locate environmental alterations [7].

BD requires the study and thinking about both technical and business needs. There are people who need to investigate technological specifics, whereas others need to know the cost-effective of BD equipment usage. Applying a BD setting will need an architectural and business method and lots of planning [5, 8].

To manage BD, data scientists are needed because there is an immense amount of data available where, in the past, there were no algorithms able to manage BD. In the past, large amount of data could not be stored. Now,
Exabyte storage and the tools needed to manipulate BD are available and not expensive. Data virtualization and efficient preservation of BD are now using cost efficient cloud storage [5].

BD technology is an essential progress track in the area of Internet science and technology. It has been broadly evaluated and progressed entirely around the globe and has been used in various areas of manufacturing as well as life [9].

There are a lot of advantages in utilizing BD technologies. BD could produce benefits like maximizing organizational productivity, informing strategic positioning, improved client services, recognizing and evolving modern products and services if utilized in an efficient way [10, 11]. Other advantages of utilizing BD is to improve marketing, automated decision making, descriptions of client behaviors, better profit on investments, quantification of dangers and market trending, understanding of commercial changing, planning and predicting, recognition of client activities beginning with click streams as well as manufacture income expanding [12, 13].

According to the knowledge of the authors, there is no single paper that gathers the basic concepts of BD which include BD definition, types of data, technologies to deal with BD, characteristics of BD, life cycle of BD, architecture of BD, applications of BD, its platforms, challenges of BD, limitations to implement BD, as well as motivations to doing BD research. Thus, this paper tries to fill in the gaps with the aim to provide the understanding of BD in a simple and easy way.

In addition to the introductory section, Section 2 presents the types and properties of BD followed by applications of BD in Section 3. Challenges in performing BD research are presented in Section 4. Finally, Section 5 presents the conclusion.

2. Types and properties of big data

There are three main categories of data within BD. The categories are structured data, semi structured data and unstructured data [1, 14].

Structured Data usually denotes data that have a described length and format. Such data includes string and dates. The majority of experts conclude that this category of data occupies about one quarter of the available data. It is frequently kept in a database and are created by machine or human. Structured data created by human include input data, like human’s names and human’s age, click-stream data game’s moves data, while structured data created by machine include sensor data (log in data to website, data selling port, and commercial data) [5].

Unstructured Data are data that are unable to keep in a particular formula. Usually, unstructured data comprises three quarter of any company’s data. Unstructured data may be located everywhere and are created by machine or human. Those created by machine include satellite images, scientific data, pictures as well as video and sensor data. Unstructured data created by humans include texts in the social media, web site and cellular phone [5, 15].

Semi Structured Data are the data that are not categorized as structured and unstructured data. This type of data do not essentially adapt as stable representation but may contain uncomplicated values [5, 15].

The 5Vs is used to describe BD [9, 16] which is also known as the characteristics of BD that relates to volume, velocity, variety, veracity and value [17, 18, 19].

Volume which is the quantity of data. This characteristic is foremost in human minds when handling BD. Many businesses have huge volumes of data that are archived in the form of logs, but do not have the ability to manage them. The benefit obtained from the ability to handle huge quantities of data to produce information is the most important desirability of BD analytics [17]. The huge volume of data can be helpful for businesses, but it affects the retrieval and analytic procedures as it is time consuming because of the calculation processes [8].

Velocity is the speed of the data that during handling and also the speed at which the data are generated. Thus, this implies the quickness that has to be considered in managing, storing and analyzing the data. Each second of each day, hundreds of hours of video are uploaded on YouTube and more than 200 million emails were sent via Gmail.

Variety refers to the range of data types and sources. The group to which BD is assigned is also an important feature that requires to be recognized by the data analysts, because BD is not usually structured data and it is not usually easy to maintain in a relational database. The difficulty of keeping and analyzing BD will maximize the difficulty of handling structured and unstructured data as 90% of created data are unstructured data.

Veracity is related to the truth of data which is important for precision in analysis. It is impossible to ensure that all data are 100% accurate when managing huge volume of data with great velocity as well as variety. The quality of the data will vary and precision of analysis depends on the data veracity of the data resource.

Value is the importance of the data importance and this is a very significant feature in BD. The possible value of BD is substantial. However, without suitable access its value cannot be exploited.

Table 1 has been provided by [20] to show the summary of 4 characteristics of BD.
Table 1. Summary of the 4V’s.

| Description                  | Volume                        | Velocity                                | Variety                                                   | Veracity                                  |
|------------------------------|-------------------------------|-----------------------------------------|-----------------------------------------------------------|-------------------------------------------|
| The amount of data created   | Exabyte, Zettabyte,           | Batch, Near/Real time streams           | Grade of structures                                        | Reliability, totality, integrity, uncertainty |
| is huge compared with        | Yottabyte, etc.               |                                         | difficulties                                              |                                           |
| standard data resources      |                               |                                         |                                                           |                                           |
| Data are being created very  |                               | Data capture from various resources     | Quality and origin of data                                |                                           |
| speedily, at a progression   |                               | like machines, people, operations both  |                                                           |                                           |
| that never ends and at which  |                               | from outside and inside the company     |                                                           |                                           |
| data is converted into       |                               |                                         |                                                           |                                           |
| perception                   |                               |                                         |                                                           |                                           |
| Data capture from various    | Maximized data resource,      | Enhanced linking                        | Mobile, social media, video,                             | Fee, requires traceability and explanation|
| resources like machines,      | higher solution, sensors,     | competitive benefit pre-calculated      | genomics, Internet of Things                              |                                           |
| people, operations both       | scalable infrastructure       | information                              |                                                           |                                           |
| from outside and inside the  |                               |                                         |                                                           |                                           |
| company                      |                               |                                         |                                                           |                                           |

The life cycle for BD consists of five phases. These phases are data collection, date cleaning, data classification, data modelling and data delivery. Data collection phase includes the gathering and storing of data from different resources. In the data cleaning phase, management of the confirmation whether there are some unwanted items within the data or missing values are performed. Data classification will classify the data according to their types either structured, semi-structured or unstructured. In the data modeling phase, analysis of data is performed and the result is the clustered data for objectives. Finally, data delivery phase involves the creation of reports based the results of the modelling phase. These phases are depicted in figure 1 [5, 21].

Figure 1. Life cycle of BD.

Several questions need to be considered in establishing any BD architecture. These questions relate to how much data will a company require to handle BD now and in the future and how will a company usually require handling data in actual time or near to actual time? Other technical issues that need to be resolved are the speed and accuracy of the data. The elements required in any BD architectural establishment is shown in figure 2 [5].

Figure 2. Big Data Architecture.
3. Applications of big data

Organizations are uncertain as to how data can be used and when to utilize the data. Applications are invented particularly to gain benefit of the exclusive features of BD especially in the fields of health care, manufacturing and city planning [5]. Figure 3 compares the different applications based on the 3V’s (variety, velocity, and volume) [22].

| Application                      | Variety | Velocity | Volume |
|----------------------------------|---------|----------|--------|
| Banking and securities           | Low     | Very high| Very high|
| Media and entertainment          | Very high| High     | Very high|
| Healthcare providers             | High    | Very high| Moderate|
| Education                        | Low     | Very low | Very low|
| Manufacturing and natural resources| High | Very high| Very high|
| Government                       | Very high| High     | Very high|
| Insurance                        | Moderate| Moderate | Moderate|
| Retail and wholesale trade       | Moderate| High     | High   |
| Transportation                   | Moderate| High     | High   |
| Energy and utilities             | Moderate| High     | High   |

Figure 3. Comparison of different data attributes in BD applications.

Other applications of BD as reported in [15] are: smart grid, e-health, Internet of Thing (IoT), public services, transportation and logistics, and political services and government monitoring.

**Smart Grid Case:** Smart grids involve the forever progressively huge size of data [23]. This is a critical area that requires real-time monitoring for almost all of its operations. This is achieved through connected devices that form the whole of the infrastructure. BD analytic helps to provide an insight like identifying deteriorating critical equipment on national grid that exhibit abnormal behavior like faulty transformers. In such cases, proactive measure and the best line of preventive and maintenance actions can be deployed thereby saving cost and optimizing operation.

**E-health:** Health related policies are among various fields that adapt and benefited from BD. Patients monitoring sensors, laboratory data, medical history of patients with different ailments are among few various sources of useful data that if utilized properly can aid into personalized medication, help policy makers to provide adaptive health care policies and it can also be utilized to reduce general hospital operational running cost and enhance service delivery.

**Internet of Things:** This is another area that benefited much from DB is IoT due to variety of interconnected objects that consume, generate and share different types of data. Objects in IoT can be anything that is capable of being connected to and be accessed online.

**Public Services:** Public services like public water systems are putting in place sensors to monitor consumption, illegal connection and leakage on pipelines in order to benefit from real time monitoring of infrastructures. This helps to reduce manpower needed for monitor facilities which resulted into timely intervention when needed and helps in rendering efficient service to public.

**Transportation and Logistics:** Transportation sector is among the most prominent areas where application of BD cannot be over emphasized. Near field communication devices like radio frequency identification enable transporters to have fleet equipped with capable sensors attached to commuting vehicles. This enable administrators to efficiently plan and manage delivery routes, access to up-to-date track record of employees, have the ability to monitor fleet in a real-time manner, and access to up-to-date pattern of passengers commuting behaviour which enables optimised planning and effective management.

**Political Facilities and Government Observing:** Several governments are mining social data to observe political trends as well as analyze community opinions. Governments may utilize BD systems to enhance the usage of scarce resources and services. For example, data from sensors placed on various public infrastructure like water pipes can be used to determine consumption rate of different regions and several zones of a city which can result into provision of required quantity of the needed resource.

Big organizations (like Yahoo, Google, and Facebook) required to develop modern tools that could permit them for storing, accessing, and analyzing massive amount of data close to real time. Platforms like MapReduce,
Hadoop, YARN, Oozie, Flume, Hive, HBase, Apache Pig, Apache Spark, Sqoop, Zoo keeper, and Big Table are modern creation of data management tools that can be used to analyze huge amount of data effectively and timely [5].

**MapReduce:** Is one of the data processing options that can be executed on Hadoop [15, 24]. MapReduce has been design to handle and schedule data processing job and cluster assignment effectively. The main advantage of MapReduce is its simplification of processing huge volume of data which is achieved through effective computing resource sharing mechanism via parallel processing procedure. The effectiveness of MapReduce is achieved through its capability for distributing task through many available mapping nodes. The quantity of maps is often concluded through the entire size of the inputs which is the entire number of blocks of the input files. Once the computation task on various distributions is accomplished, extra task named as “reduce” then gathers all the processed results to produce a complete processed solution. This approach facilitates better load balancing, maximizes the quantity of reduces, maximizes load equalization and minimizes the number of breakdowns [25]. Figure 4 shows data flow within MapReduce [5].

![Figure 4. Data flow in MapReduce.](image)

**Big Table:** This was created by Google to act as a distributed saving system that aimed to process extremely huge amount of structured data using secure servers. Data is organized as a table that has tuples and attributes. Big table differs from classical relational database in various ways. It is a spare, distributed, and permanently multi-dimensional saving map [5].

**Hadoop:** There are different stories about the name of Hadoop. [26] stated that Hadoo is an acronym name stand for **Highly Archived Distributed Object Oriented Programming.** [27] stated that Hadoop is an elephant in the room. It is not an acronym but a yellow elephant toy for the creator’s son of it which is the Google’s engineer Doug Cutting. However, Apache Hadoop is a hug scalable storage platform created to manage huge amount of data sets through hundreds and thousands of calculating nodes that work concurrently [2, 17, 28]. Hadoop is also considered as an Apache handled software structure resulting from Big table and MapReduce. Hadoop permits functions to be established on MapReduce to execute big clusters of productized hardware. Hadoop is the basis of the calculating architecture providing Yahoo!’s company. Hadoop is intended to simultaneous process data through computation nodes that speed up the calculation and minimize latency. Two main elements in Hadoop are (1) huge scalable distributed file system in providing petabytes of data and (2) huge mountable MapReduce that calculates outcomes in batch. Figure 5 illustrates how to map Hadoop clusters into hardware [5].

![Figure 5. Mapping Hadoop clusters into hardware.](image)
4. Challenges in big data research

**Big Data Mining:** BD mining presents a lot of desirable chances but with immense difficulties. The complexities depend on various stages involving data taken, storage, seeking, sharing, analysis, administration and visualization [15].

**Big Data Management:** BD management aims to provide a dependable clean data through various means of data gathering from huge volumes of different types of data sources like company, government as well as private/public sectors. This is achieved by various processing tasks that include preprocessing, processing and other related activities such as encrypting the data for security, confidentiality and dependability [26]. Certainly, suitable data management is the basis for BD analytics [15].

**Big Data Recovery and Storage:** Storage in BD is accomplished via virtualization where it processes huge sets of data from sensors, media, videos, transaction data from e-businesses, mobile signal coordinates. Many corporations manage data in huge volume through utilizing instruments such as NoSQL, Apache Drill, Horton Works, SAMOA, IKANOW, Hadoop, MapReduce, and Grid Gain [15]. Large volume storage facilities and faster I/O speeds enables improvement in working with BD. Thus, access to data should be quick and simple for on-time analysis. Previously, continuing data were replaced through utilizing Hard Disk Drive (HDD). The well-known major drawback of HDD is having slower input/output performance. Improvements in storage devices like Solid State Drive (SSD) can minimize the problems but they are not been completely utilized. HDDs are gradually being replaced by SSDs, and other improvements like pulse-code modulation which are also on the increase [22].

**Big Data Processing:** BD processing analyses the huge volume of BD in petabyte, exabyte, and zettabyte depending on whichever batch management or batch management is best [26].

**Data Visualization:** The major aim of data visualization is to present the data efficiently and sufficiently through utilizing several charts. Data visualization poses a challenge for BD applications because the huge volume and dimension of the data. Thus, there will be a need to re-test the method under which the BD is pictured. Structure and usefulness of presented data are of paramount importance to permit demonstration of knowledge which is unseen in non-simple large-scale data sets. The structured data organized in tables and its associated characteristics are necessary for informative analysis [22].

**Data Transmission:** Once the communication foundation is extremely big, the system's data transfer capability is bound and blocked in a cloud circulated framework. Here cloud improvement is substituted via a cloud data feed as its improved form [22].

**Big Data Security:** It is a challenge to ensure guaranteed safety and security of big data. This is attributed to many factors like incompetent instruments, public and private databases. In distributed programming structures, the safety challenge begins when huge amounts of personal information are kept in a database which is not encoded in standard form. Saving the data in the hand of disgruntled and unreliable persons add to the extra complexity of data security. The challenge of data security also surfaced when migrating or updating from similar and/or different data-specific instruments. Occasionally, data thieves and system thieves gather a publicly accessible BD collection, copy it and keep it in a device like a USB drive, hard disk or laptop. Therefore, when keeping of data is maximized from one level to a multi-storage levels, the safety level should also be maximized [26].

**Data Curation:** This area specifically contains several sub-areas, like validation, documentation, supervision, security, recovery and demonstration. The existing database management tools cannot manage BD. Data warehouse and data marts have been used to manage big data sets in a suitably structured approach. These methods follow data frameworks which are built on structured query language. These days NoSQL is utilized in BD due to the four Vs of BD [22].

**Big Data Cleaning:** This challenge involves five phases (cleaning, aggregation, encoding, storage and access) which are not modern and are used in conventional information handling. The challenge in BD is how to handle the difficulties of BD’s nature (velocity, volume, variety, veracity and value) and operate it in a distributed situation with a combination of functions. Nevertheless, information resources may include noises, errors or incomplete data. The issue is how to clean large data and how to resolve which data is countable, and which data is beneficial [15].

**Big Data Aggregation:** This issue is related to synchronizing external data resources and distributed BD platforms (involving applications, repositories, sensors, networks) with the internal infrastructure of a company. Furthermore, it is not enough to analyze the data created inside companies. In mining important insight and information, it is necessary to move towards an advance of not collecting only internally generated data, but required tools should be put in place to collect both internal data and external data resources. External data could contain third-party resources, information about market fluctuations, weather predicting and traffic conditions, data from social networks, customer comments and citizen feedback [15].
**Big Data Imbalance:** The issue in classifying an imbalanced data set has received a lot of attention. Real world applications with diverse distributions can be categorized into two major groups. The first is the under-presented group which is characterized with insignificant number of data points (this type is also known as minority or positive group). The second group has significant number of data points (also known as common or negative group). The identification of positive group has significant importance in many areas like medical diagnosis, software defects detection, finances, drug discovery or bioinformatics. The traditional Machine Learning (ML) techniques cannot be implemented on imbalanced data sets. This is because the prototype building is founded on global privilege measures which by default favours the majority instances group thereby disregarding the significance of the minority group [15].

**Big Data Analytical:** BD brings a lot of challenges on how to extract meaning out of this large, ever-increasing voluminous data. For example, data analysis allows a company to obtain important vision and observe the patterns that may positively or negatively influence their businesses. Other data-driven applications require additional real-time analysis, like social networks, biomedicine, astronomy, and intelligent transport systems. Therefore, advanced algorithms and efficient approaches of data mining are required to obtain correct outcomes, to control the changes in different areas and be able to have future predictions with real-time responsiveness. One of the challenges is how to guarantee the timeliness of responses when the data volume is huge. Examples of the complexities observed when performing existing analytical solutions are ML, deep learning, incremental approaches, as well as granular computing [15]. The main issue of data analytic found with BD is related with the volume of data. Timeliness is the highest priority for some BD applications. The main test for BD applications is to guarantee timeliness of responses when the data being processed is huge [22].

**Big Data Machine Learning:** The primary aim of ML is to discover knowledge from either organized or unorganized data. ML is presently serving as the backbone of many applications that relies and produce part of big data composition, ranging from search engines, recognition systems, aeronautics, and military to mention a few [15].

BD is an innovation that will fundamentally change the method that information is grouping, keeping, monitoring, and spent by users which in turn will change the method of doing work. Several of the motivations in doing BD research are [22]:

**Changing from Classical Relational Database Management System (RDBMS):** This system is utilized by several enterprise information technology corporations and is still used now by many information technology enterprises. Today, data are unstructured and non-clustered and NoSQL keeps all the data with no clustering and describes them into the framework opposite to RDBMS which keeps data in suitable structures or tables.

**Managing Unstructured Data:** BD has the capability to manage structured as well as unstructured data. Along with data variety characteristics, BD is about text or numbers (alphanumeric fields), and unstructured data. Therefore, by utilizing NoSQL, BD can manage unstructured data.

**Real Time Data Processing:** In future, information systems will need an ability to manage increasingly huge volumes of data, where the velocity of BD created is presented. An expression "near real time", is often utilized with existing time of information systems, but it is not suitable. Real time data management involves the facility of managing online data or sensor information as they are created.

**Most Data are either User or Machine Created:** Earlier, most data were created internally in the firewall of an enterprise. However, present data are created either through end users or machine created, which are external to the bounds of the firewall of the enterprise.

There are many limitations in BD and some of these limitations can be summarized as follows [29]: firstly, the needed data are not always available because of: (i) data are simply not available; (ii) there is trouble with the holding phases; and (iii) various data platforms are shown to be interoperable. Secondly, the main core of BD is pattern recognition. Result from pattern analyzing is significant because it will demonstrate the problem of security, risk, and types of crime. Present approaches of data mining will not be able to handle BD. Lastly, BD can be used to predicate new data because it is established on old data which consist of previous patterns.

5. **Conclusion**

It can be concluded that BD means any quantity of data even if the data is structured, unstructured, or semi-structured that cannot fit into a processing system. Thus, BD will need special tools and technologies to handle it and can be characterized by the term 5Vs. If a company needs to model its BD and gain benefits from it, the company needs to design the architecture for its BD which will require answers to questions related to the nature of the company. BD has many applications fields in life and has many platforms to handle it. In reality, BD has many challenges and limitations to implement it, as well as there are several motivations in performing BD research.
References

[1] José T and Juan R 2018 Data learning from big data Statistics and Probability Letters 136 15-19
[2] Mitchell I, Locke M, Wilson M and Fuller A 2012 The white book of big data (UK: Fujitsu Services Ltd.)
[3] Lisa A 2016 Big data marketing (New Jersey: John Wiley & Sons, Inc.)
[4] Bernard M 2016 Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results (New Jersey: John Wiley & Sons, Inc.)
[5] Judith H, Alan N, Fern H and Marcia K 2013 Big data for dummies (New Jersey: John Wiley & Sons, Inc.)
[6] Maria B, Liyana S and Elaheh Y 2019 Big data adoption: state of the art and research challenges Information Processing and Management 56
[7] Alessandro M, Sally D, Maureen M, Lee Q, David W, Lyndon S and Ana C 2018 Big data, big decisions: the impact of big data on board level decision making Journal of Business Research 93 67–78
[8] Michele I, Elio M, Giuseppe M, Mario M and Carlo Z 2020 Fast and effective big data exploration by clustering Future Generation Computer Systems 102 84-94
[9] Yinghao Y, Meilin W, Shuhong Y, Jarvis J and Qing L 2019 Big data processing framework for manufacturing Procedia CIRP 83 661-64
[10] Jaime C, Pankaj S, Unai G, Erkki J and David B 2017 A big data analytical architecture for the Asset Management Procedia CIRP 64 369-74
[11] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C and Hung B 2012 A big data: the next frontier for innovation, competition, and Productivity (McKinsey Global Institute)
[12] Seref S and Duygu S 2013 Big data: a review Proc. Int. Conf. on Collaboration Technologies and Systems (CTS) (San Diego: CA/ USA IEEE) p 42
[13] Philip R 2011 Big Data Analytics TDWI Best Practices Report
[14] Karim M 2019 State of the art in big data applications in microgrid: a review Advanced Engineering Informatics 42
[15] Ahmed O, Fatim-Zahra B, Ayoub A and Samir B 2018 Big data technologies: a survey Journal of King Saud University-Computer and Information Sciences 30 431-48
[16] Ada B, Devis B, Valeria A, Massimiliano G and Alessandro M 2019 A relevance-based approach for big data exploration Future Generation Computer Systems 101, 51–69
[17] Ishwarappa K and Anuradha J 2015 A brief introduction on big data 5 vs characteristics and hadoop technology Procedia Computer Science 48 319-24
[18] Jean-Louis M and Soraya S 2016 Big data, open data and data development (ISTE Ltd and John Wiley & Sons, Inc.)
[19] Abdul Khalique A, Vlad K and Michael B 2017 Addressing barriers to big data Business Horizons 60 285-92
[20] https://courses.cognitiveclass.ai/courses/coursev1:BigDataUniversity+BD0101EN+2016_T2/courseware/407a9f86565c44189740699636b4fb85/12eab34ec2184f689954e4d06566eef4a32
[21] Archennaa J and Mary E 2015 A survey of big data analytics in healthcare and government Procedia Computer Science 50 408-13
[22] Kirtida N and Abhijit J 2017 Role of big data in various sectors Proc. Int. Conf. on IoT in Social, Mobile, Analytics and Cloud (I-SMAC) (Palladam/India IEEE) p 117
[23] Tom W, Nalini J, Peter F and Joshua T 2019 A big data platform for smart meter data analytics Computers in Industry 105 250–59
[24] https://www.ibm.com/analytics/us/en/technology/hadoop/mapreduce/#what-is-mapreduce
[25] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
[26] Saraladevi B, Fazhanurajam N, Victer P, Saleem M S and Dhavachelvan P 2015 Big data and hadoop—a study in security perspective Procedia Computer Science 50 596-601
[27] https://www.sas.com/en_us/insights/big-data/hadoop.html
[28] https://www.ibm.com/analytics/us/en/technology/hadoop
[29] Dennis B, Erik S and Bart S 2017 Big data and security policies: towards a framework for regulating the phases of analytics and use of big data Computers Law & Security Review 33 309-23