Journal of Engineering Sciences

Vol 13, Issue 06, JUNE /2022

ISSN NO:0377-9254

A Comprehensive Survey of Big Data in the Age of AI

[1]T. Aditya Sai Srinivas, [2]Anday Shanthi Priya
[3]Boyapalli Shanmukha Priya
[1] ASSISTANT PROFESSOR
[2, 3]B.TECH, G.PULLAIAH COLLEGE OF ENGINEERING AND TECHNOLOGY

Abstract

AI has radically altered the way we live. A solitary main reason AI works is the rapid expansion of data. Data Processing is advancing in lockstep with AI. Many aspects of society have more advantages from the use of data processing, and it appears that we are all now dependent on it. Data processing, and also it, has its downsides. Those articles will give an outlook of the current data processing technology, as well as its applications and possible flaws. Finally, we touch on the future of data processing in the context of AI and IoT.

Keywords: Big Data, Artificial Intelligence (AI)

1) INTRODUCTION

Mobile devices and the Internet of Things have led to an explosion in data over the last few decades. Is there any significance to these stats? When dealing with large amounts of data, how do you make them shine? Or would you rather just throw it away? Cache, cleaning, processing, and decision-making research have all been presented in recent years as new big data processing technologies. Big data has a profound impact on every facet of our lives[1].

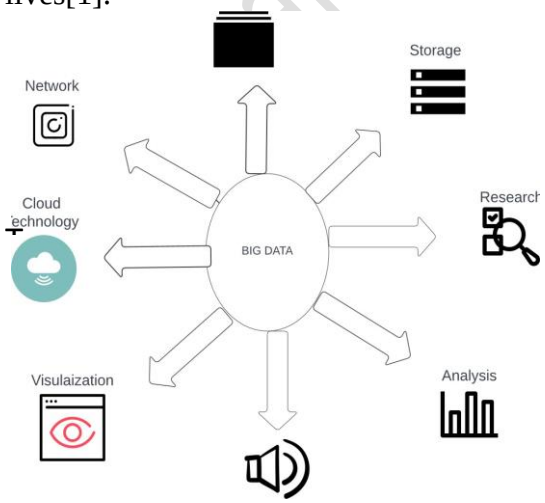

Fig. 1.Big data

The most well-known use of big data is in network applications, in which enterprises can use social media and browser data to build customer prediction models and learn more about their habits, tendencies, and hobbies. Big Data isn't just about collecting data from sensors; it can also come from humans, texts, images, and other sources [2]. Big Data has had a significant impact on technology and computing. Data Processing refers to the more collection, processing, presentation of huge amounts of data that arrive at more speeds in types of formats. The data processing in treatment can decode the total chromosome in a minute. Finds new treatments and understands disease patterns [3]. Connected with sports events now a day's use big data technology. Moreover using the video check to follow the football positioning or player performance and after that get report for better level competition.

Our lives are increasingly influenced by data processing as a result of AI. Data processing is a relatively new

phenomenon, which we'll discuss in this piece [4].

2) BIG DATA TECHNIQUES:

A. Data Recording

Now days in the globe the usage of technology is increase the more data. It's not just limited to a few terabytes of storage on a hard drive. And most computer storage devices are unable to store that much data. “Data stored in disc arrays must be accessible quickly and reliably. Multiple disc arrays are used simultaneously in distributed storage to store data. Because of their design pattern and data storage mode, relational databases were once limited to a single machine [5]. That is to say, no matter how much data is generated, only one machine will be required to store and manage it. All data is stored on each node, even if it is clustered. The storage capacity of each machine is limited to a few terabytes [6]. In addition, the speed of data retrieval decreases as the magnitude and document dimension of a database increase. Many of the most popular databases have come up with solutions to this problem.

1) Read and write separation:

The goal of read/write separation is to reduce the load on a single database server by distributing read/write operations among several database servers at once. To keep reading and writing operations separate, a master-slave database is employed. It is possible to have a master and a slave database in the same system [7].

2) Distributed database:

The ever-increasing demands on a single powerful server cannot be met. Despite the fact that only one server that runs and divide into two or more3, they are still unable to meet the increasing demands of a growing business. Tables are spread across many databases in a distributed database, which lets multiple servers handle requests at the same time[8].

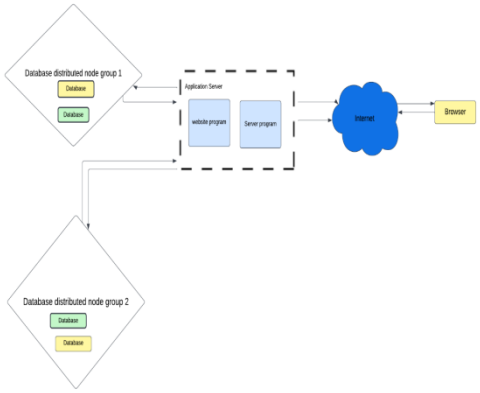

Fig: 2 Distributed deployment database

3) APPLICATION SERVICE AND DATA SERVICE SEPARATION:

Dividing the application servers from database has the more advantages of allowing the bottom structure to be tailored to the specific features for each server separating the database server from the application server is a good idea because the database server uses a lot more disc space. This way, any issues with one server won't affect the other [9].

B. Data Clean:

It's common for data cleaning to be overlooked in favour of other data processing steps. The data cleaning challenges is important and also to increase the standard result directly and also the models of the overall result. Theoretically, there are several stages to data cleaning [10]:

1) Data pre-processing:

Bring the data into the tools for analysis. MATLAB and SPSS are widely used for data analysis. Data annotating and field descriptions help you to better understand the information represented by a data

visualisation. Some of the data can also be visualised. An outlier is a data issue that we can understand intuitively [11].

2) Missing value cleaning:

If a field has a high rate of missing values, there are a variety of ways to clean it up. To complete the information that is missing, you can use pre-existing data or make deductions or estimates based on your own professional experience [12]. Removing or supplementing fields with low importance and a high missing rate is possible. Low-value data, as well as data with a low rate of missing, can be estimated. It's possible to remove fields that aren't important and have a lot of missing data [13].

3) Format content cleaning:

Structure of the metadata that is collected usually consistent and reasonable. Manual data, on the other hand, is prone to errors. For instance, the mobile phone number is entered as an ID card number, the date is incorrectly entered, and the gender representation is incorrectly entered. Proofreading is the most time-consuming part of the process. It includes both automated and manual steps [14].

4) Logical error data cleaning:

Many logical errors are caused by poor data verification because most data is entered manually. On the ID card, for example, the age is incorrectly listed as 200. In spite of how rare logical errors are, they still require more time to clean up data because they can't be missed [15].

5) Remove unnecessary data:

The more data you have, the better your models will be, but the processing time will be longer. Data processing can be speeded up and model interference reduced by tailoring unnecessary data. This makes it difficult for data analysts to determine if a piece of information is really needed. There must be a theory to back up what you're doing before you can remove any data from your computer [16].

6) Data Validation:

With multiple data sources, there is the possibility of unpredictability, which is difficult to detect but has an impact on the model's ability to predict. Essentially, a well-executed data acquisition strategy can avoid this issue [17].

C. Data Analysis:

Big data technology necessitates the use of data analysis, also referred to as data mining. In the decision-making process for the completed Hu data, the data analysis result is the most important information available. A lot of progress has been made recently in the field of data analysis. Data analysis methods are being studied more and more by academics. Here are a few methods commonly used for analysing data [18].

1) Classification:

The goal of semi-supervised learning techniques for classification is to train a model that returns one of multiple possible classes for each example using labelled and unlabelled datasets. We consider the best-performing techniques in a survey that focuses on labelling data, which are summarised in the figure below [19].

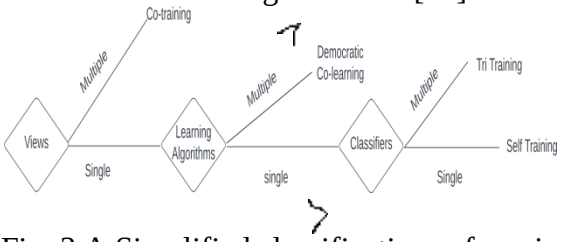

Fig: 3 A Simplified classifications of semi-supervised learning techniques

The performance results are comparable whether transudative or inductive learning is used. [3.1.1]. the majority of the time, the data is complex, and illogical data frequently has an effect on the analysis. The simplest class of semi-supervised learning techniques trains a single model

on a single set of features using a single learning algorithm, and also classifying data and then mining it for specific categories is a straightforward processing method [20]. The following class trains multiple classifiers by performing multiple samples of the training data and training a model for each sample. This classification and processing method can take advantage of the data's inherent characteristics [21]. Whereas each model is updated iteratively, the remaining models make predictions using the unlabelled data, and iteration ends when no model is changed. Finally, the unlabelled models are labelled using majority voting, which requires that at least two models agree. This process is repeated until no additional data is added to the classifier's training data [22]. The final class makes use of multiple views, each of which is a subset of features that is conditionally independent of the given class. According to the survey, these algorithms achieve comparable transudative or inductive precision.

Regression:

Semi-supervised learning for regression has received relatively little research, with the goal of training a model that predicts a real number. Regression is a common technique for analysing large amounts of data. It is a favourite amongst data analysts for its mathematical push ability and rigour. Data in a dataset must be linked together, which means there must be independent and dependent variables[23]. Create a mathematical model of how the data is spread out based on what you know or what you think. This can be done with regression analysis, and it can help find the model parameters that best match the data. It is possible to use the mathematical evaluation method once we have the model in place. The model's reliability must meet a set of standards before it can be used for forecasting and analysis[24].

2) Clustering:

It is possible to achieve known clustering through certain rules, and this is what we mean by "clustering." It categorises data that is similar in nature. They're exactly the same.And also, the data to be classified in a types of ways. According to experiments, clustering algorithmslike Lloyd's can be unsupervised classified according to the number of classes expected by data analysts [25].

3) Statistical Description:

Tables and icons are used to differentiate and also analysing the data in a statistical description. That is both intuitive and effective [26].

D. Data Visualization:

There is a lot more to data visualisation than simply looking at a spreadsheet. Good data visualisation takes into account the data's context, distribution, and even colour expression. Using a histogram or pie chart, for example, we can depict the employment rate in various provinces of the country [27]. If the background is a map with different colour spectra for each province, the data observer can get the data more intuitively and quickly. When making a visual representation of data, things like how accurate the data needs to be, who the target audience is for data analysts, and other things need to be taken into account [28].

E. Decision making:

It is important to note that all previous steps lead to data decision-making and their quality influences the outcome. It has become increasingly common in recent years for decision theory to focus primarily on empirical decision making. Researchers have come up with a variety of methods for determining whether or not a study's findings are accurate [29]. When it comes to AI's impact on decision-making, this article focuses on methodological shifts. For the most part, humans are incapable of

comprehending and digesting today's vast amounts of data in the data processing era. Human observation of large amounts of high-dimensional data, on the other hand, consistently yields suboptimal outcomes [29]. Some researchers have proposed statistical machine learning methods, which is encouraging. Data can be fitted to and future events predicted using a probability model in most cases. Making a decision based on the prediction is possible. This is how the artificial intelligence decision model works. There are numerous branches of machine learning, all of which have produced stunning results. An example is provided here [30].

1) Classification Tree:

The Classification tree is most commonly used methods for classifying the data. Using a decision tree, decision-makers are walked through the process of making a final decision step by step. A well-known example is the chess grandmaster's dark blue. This successful case influences AlphaGo's in-depth learning version. Using a decision tree is a straightforward process[31]. We can use chess pieces to create branches just like in the game. It is possible to determine the optimal result of the branch by simulating it. Due to current computer memory storage technology, such algorithms cannot run for long periods of time. The traditional of optimization is pruning, which means ignoring branches with low winning rates[32]. Decision trees are strong, interpretive, and stable algorithms despite their high computational and pruning requirements[33].

2) Deep learning:

Deep learning has had some notable successes, as I'm sure you've all heard. The prediction branch has been a popular alpha go-to in recent years, but with more machines, the decision tree it has can go deeper. As a result, we're in a better position to make a decision. So how does he determine the value of a node in a decision tree? When two chessboards are compared, how can Alpha Go tell which one is better? This appears to be a simple problem, but it's actually quite difficult. Consider it from a different perspective[34]. Inexperienced players may have difficulty distinguishing between the two sets of boards, but can an expert tell? Because of the players' wealth of knowledge, this is tolerable. Chess books are easy to remember for an experienced player. Are historical chess tree simulations and analysis possible for this process? This can only be done with deep learning. Aside from that, it's capable of improving and optimising its own decisions over time[35].

F. Pre-processing:

The quality of the data that is analysed is critical for good decision-making. Prior to further analysis, data must be pre-processed by removing inconsistencies, incompleteness, and numerous errors. It sets the stage for further processing and analysis by preparing the data for it. It's possible to achieve the pre-processing section goal by following these steps[36].

1) **Data cleansing**: A process of removing errors, inconsistencies, and incompleteness from data
2) **Data transformation:** As a result, additional operations such as aggregation or transformation are performed. It has a significant impact on the next steps that are taken[37].
3) **Data integration:** It gathers distributed data from a variety of data sources.
4) Data transmission: It is the transfer of data from one digital device to another device.

The following sub-sections provide additional information about several pre-processing steps:

1) **Data Transmission**: It transfers raw data to a data storage location [38].

2) **Data Cleansing: It** is process of fixing incorrect, incomplete, duplicate or otherwise

Erroneous data in data set common feature is tool in data enrichment. Maletic and Marcus considered five stages in achieving clean data:

1. Identifying the various types of errors
2. detecting instances of errors
3. correct instances and types of errors
4. Update the data input procedure to eliminate any potential errors.
5. Limitations, formats, and rationalities of data should all be checked [39].

Data cleansing is an important and necessary step in the data analysis process. In a nutshell, there are two major issues with the data cleansing step:

i) Data are imprecise

ii) Data is incomplete (parts of the dataset are missing), and we should address these issues as much as possible.

3) **Stream Processing:** The processing of stream data has been a challenge for researchers in the Big Data field. With traditional batch processing, the stream requirements are completely different [41].There is a number of open research topics in the stream processing section, which are listed below:

1. **Data Mobility**: It refers to the number of steps required to obtain the desired outcome [42].

2. **Data Division or Partitioning:** Data is partitioned using the algorithms. In summary, partitioning strategies should be employed to improve data parallelism.

3. **Data Availability:** We should propose a method that ensures data availability in the event of a failure.

4. **Query Processing:** We should propose a query processor for distributed data processing that is efficient and takes data streams into account. Doing deterministic processing (always getting the same answer) is one option, while non-deterministic processing (the output depends on the current situation) is another [43].

5. **Data Storage:** A further open Big Data research question is how to store data for later use.

6. **Stream Imperfections:** Methods for dealing with data stream flaws such as delayed or out-of-order messages [44].

BIG DATA APPLICATION

The statistical analysis of business data is the goal of big data applications, which are developed to aid systems and businesses alike. Data granularity is decreasing as data content and formats become more diverse. Distributed storage, computation, and streaming are now available. Each industry investigates more application scenarios based on multiple or cross-industry data sources in order to achieve individual-oriented decision-making and application timeliness. A big data application is distinct from a traditional data application in terms of the data format, processing technology, and application form [45].

A. Telecom

As is well-known in the telecom industry, the mobile phone users of a single operator generate enough data to fill a PB's worth of storage capacity. The telecom industry has been collecting data for decades in order to improve network performance and provide better customer service thanks to the use of information technology in the

industry. It's possible to use only a fraction of the data resources under traditional processing technology. There should be a greater emphasis on data awareness, business models that take advantage of data, and an effort to understand data's true value for telecom operators. (Data as a Server and Application as a Service) are the two most common patterns (Analytics as a Service). De-identified data can be sold directly to the public via open APIs or public data. To generate external revenue and realise the realisation of data resources, the AaaS model frequently collaborates with a third-party company to provide general information and services to government, enterprise, or industry customers after the desensitisation of data resources [46].

B. Traffic

Real-time traffic data is a common type of big data. The use of big data in transportation is also well-established and productive.

1) For travel:

Integrate information about travel services using big data. Additionally, real-time road and flight data can be provided via this system.

2) For logistics:

Forecasting logistics markets and streamlining distribution systems can both benefit from logistics data. Even the dynamics of socioeconomic change can be gleaned from it.

C. Medical

Genomic, clinical trial, environmental, resident, and health management data are just some of the many types of big data. Clinical protocols and decision-making can be guided by this data. Epidemiological data can be used to assess and predict disease risk at the same time. Comprehensive research and on-going error correction allow this system to provide the most accurate diagnoses and treatments possible. By analysing unstructured data, big data analytics can improve clinical decision-making.

3) BIG DATA PROBLEM:

A. Challenges of Data Storage

The amount of data generated on a daily basis by the Internet is increasing at an exponential rate as more mobile devices are connected to it and as more people use it. Massive amounts of data are too much for the original infrastructure, such as storage and transmission. In order to lower the cost of discs and increase their storage capacity, more and more academics have begun to invest in storage research. The recent installation of optical fibre in network data transmission equipment has greatly accelerated network data transfer speeds. Techniques like local cache and local cache are used by the industry to speed up the process of getting data.

B. Process Speed Challenge

Moore's law states that a CPU's processing speed should increase by two times every two years. Industrial production technology has reached a point where Moore's Law is no longer applicable. Because of this, the amount of heat generated by high-density calculations will rise dramatically, posing new problems for materials. Consequently, an increasing number of academics are devoting time and resources to investigating CPU speed. CPUs with multiple cores and a distributed infrastructure are popular right now. That is, there are multiple chips on the current host. It's not uncommon for workstations to have two, four, or even eight cores. CPUs have a finite number of cores and can't keep up with the exponential growth of data. This is how it works now: The industry can process data on hundreds of hosts at the same time with the help of distributed infrastructure and distributed algorithms, then summarise the results.

For the purpose of proving the well-known cap theorem, we need three things: consistency, availability, and partition. There is usually a trade-off in a distributed system because of the cap theorem. To get the most out of the distribution, users must take responsibility for their own needs and be willing to make reasonable compromises.

C. Personal Privacy Issue

Big data not only provides benefits and opportunities, but it also keeps us open to the general public. The "double-edged sword" aspect of big data is perfectly exemplified by the "Prism Gate" incident in the United States. Deep mining increases the value of data while also revealing more personal information about network users. Deep mining There are two ways in which big data privacy leakage occurs:

1) Consumption guidance:

In order to predict consumption patterns, you can search for consumer records and guide your own consumption. Personal services appear to be offered whenever and wherever the customer desires them to be. As a result, the privacy of consumers is being violated by using this data without their permission.

2) Identity information disclosure:

Every day, people's locations and steps are revealed by their mobile phones. Browser history reveals one's thoughts and feelings on a daily basis. Through the use of data mining techniques, information about an individual's identity can be discovered.

4) BIG DATA FUTURE:

Big data has a positive impact on our lives and makes them more convenient in the era of artificial intelligence. In the future, big data will be used with applications in a wide range of ind3ustries to solve real-world problems.

A. Collaborate with AI

As neural network algorithms are optimised, processors become faster, allowing for more complex network training to be performed. Thereafter, an extremely deep neural network is built. Data-driven, complex neural network learning systems, such as deep learning, are evolving traditional neural networks. The amount of data on the Internet has exploded in recent years. Big data provides the processing power, while AI provides the scenarios for putting that power to use. The rapid growth of massive data and big data applications necessitates the use of AI technology to boost processing capacity when the data is spread across multiple systems and businesses. Big data is fuelling a new era of rapid AI development. As a result, it is important that AI and big data work more closely together.

B. Collaborate with The IoT

In the Internet of Things (IoT), there are already many technologies, such as communications and big-data analytics, in place. Its primary function is data transmission. Traditional communication networks have been expanded with the Internet of Things (IoT). The Internet of Things (IoT) has access to a vast array of devices, which greatly enhances the network information and data sources. Scale and industrial impact are difficult to achieve with simple and partial IoT applications because they use simple data types and small data volumes. IoT's shortcomings can be compensated for by big data technology and data mining capabilities. Intelligent analysis is made possible by big data and AI-related technologies, such as machine learning. More and more data is being collected via the Internet of Things, and as the technology improves, so does the demand

for IoT and big data technology integration.

5) CONCLUSION

The importance of data is only going to rise in the age of artificial intelligence (AI). Keeping up with the advancement of science and technology in the era of artificial intelligence necessitates a thorough understanding of big data, its algorithms, and its benefits. A brief introduction to big data methods and current issues was provided in this paper, and we examined the use of big data in social reality. Big data's future development is also worth keeping an eye on.

REFERENCES:

[1]T. Aditya Sai Srinivas, S. Ramasubbareddy, and K. Govinda, "Loan Default Prediction Using Machine Learning Techniques," in Innovations in Computer Science and Engineering, Springer, 2022, pp. 529–535.

[2]S. S., R. Somula, B. Parvathala, S. Kolli, S. Pulipati, and A. S. S. T., "SOA-EACR: Seagull optimization algorithm based energy aware cluster routing protocol for wireless sensor networks in the livestock industry," Sustainable Computing: Informatics and Systems, vol. 33, p. 100645, 2022, doi: https://doi.org/10.1016/j.suscom.2021.100645.

[3]T. Aditya Sai Srinivas, S. Ramasubbareddy, A. Sharma, and K. Govinda, "Optimal Energy Distribution in Smart Grid," in Intelligent Data Engineering and Analytics, 2021, pp. 383–391.

[4]T. A. S. Srinivas, S. Ramasubbareddy, K. Govinda, and C. S. P. Kumar, "Storage Optimization Using File Compression Techniques for Big Data.," in FICTA (2), 2020, pp. 409–416.

[5]T. A. S. Srinivas, S. Ramasubbareddy, A. Sharma, and K. Govinda, "Optimal Energy Distribution in Smart Grid.," in FICTA (2), 2020, pp. 383–391.

[6]S. Sankar, R. Somula, B. Parvathala, S. Kolli, S. Pulipati, and others, "SOA-EACR: Seagull optimization algorithm based energy aware cluster routing protocol for wireless sensor networks in the livestock industry," Sustainable Computing: Informatics and Systems, vol. 33, p. 100645, 2022.

[7]T. Aditya Sai Srinivas, R. Somula, K. Aravind, and S. S. Manivannan, "Pattern Prediction Using Binary Trees," in Innovations in Computer Science and Engineering, Springer, 2021, pp. 43–52.

[8]T. Aditya Sai Srinivas, R. Somula, and K. Govinda, "Privacy and security in Aadhaar," in Smart Intelligent Computing and Applications, Springer, 2020, pp. 405–410.

[9]T. A. S. Srinivas and S. S. M. Manivannan, "Preventing collaborative black hole attack in IoT construction using a CBHA--AODV routing protocol," International Journal of Grid and High Performance Computing (IJGHPC), vol. 12, no. 2, pp. 25–46, 2020.

[10]T. Srinivas and S. S. Manivannan, "Black Hole and Selective Forwarding Attack Detection and Prevention in IoT in Health Care Sector: Hybrid meta-heuristic-based shortest path routing," Journal of Ambient Intelligence and Smart Environments, no. Preprint, pp. 1–24, 2021.

[11]A. S. S. Thuluva, M. S. Somanathan, R. Somula, S. Sennan, and D. Burgos, "Secure and efficient transmission of data based on Caesar Cipher Algorithm for Sybil attack in IoT," EURASIP Journal on Advances in Signal Processing, vol. 2021, no. 1, pp. 1–23, 2021.

[12]S. Ramasubbareddy, T. Aditya Sai Srinivas, K. Govinda, and S. S. Manivannan, "Crime prediction system,"

in Innovations in Computer Science and Engineering, Springer, 2020, pp. 127–134.
[13]S. Ramasubbareddy, E. Swetha, A. K. Luhach, and T. A. S. Srinivas, "A Multi-Objective Genetic Algorithm-Based Resource Scheduling in Mobile Cloud Computing," International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), vol. 15, no. 3, pp. 58–73, 2021.
[14]D. Opresnik and M. Taisch, "The value of big data in servitization," Int J Prod Econ, vol. 165, pp. 174–184, 2015.
[15]Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," Information Fusion, vol. 42, pp. 146–157, 2018.
[16]L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," Data Sci J, vol. 14, 2015.
[17]D. Mourtzis, E. Vlachou, and N. Milas, "Industrial big data as a result of IoT adoption in manufacturing," Procedia cirp, vol. 55, pp. 290–295, 2016.
[18]B. T. Hazen, J. B. Skipper, C. A. Boone, and R. R. Hill, "Back in business: Operations research in support of big data analytics for operations and supply chain management," Annals of Operations Research, vol. 270, no. 1, pp. 201–211, 2018.
[19]F. Mazzocchi, "Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science," EMBO Rep, vol. 16, no. 10, pp. 1250–1255, 2015.
[20]V. N. Gudivada, R. Baeza-Yates, and V. v Raghavan, "Big data: Promises and problems," Computer (Long Beach Calif), vol. 48, no. 03, pp. 20–23, 2015.
[21]M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," J Parallel Distrib Comput, vol. 79, pp. 3–15, 2015.
[22]Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 1891–1899, 2017.
[23]A. de Mauro,M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," Library Review, 2016.
[24]K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," Renewable and Sustainable Energy Reviews, vol. 56, pp. 215–225, 2016.
[25]X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," Big data research, vol. 2, no. 2, pp. 59–64, 2015.