Big Data Security Challenges and Preventive Solutions

Nirmal Kumar Gupta and Mukesh Kumar Rohil

Abstract Big data has opened the possibility of making great advancements in many scientific disciplines and has become a very interesting topic in academic world and in industry. It has also given contributions to innovation, improvements in productivity and competitiveness. However, at present, there are various security risks involved in the process of collection, storage and use. The leakage of privacy caused by big data poses serious problems for the users; also the incorrect or false big data may lead to wrong or invalid analysis of results. The presented work analyzes the technical challenges of implementing big data security and privacy protection, and describes some key solutions to address the issues related with big data security and privacy.

Keywords Big data · Big data analysis · Big data security · Privacy protection

1 Introduction

In today's world, a large number of people share their social information and behavior using the internet and it has led to the explosion of data generated. The constant advance of technologies has allowed an "explosive" growth in the amount of data generated from different sources, for example social networks, mobile devices, sensors, X-ray machines, telescopes, space probes, applications logs, climate predictions, geo-positioning systems and, in general terms, everything that can be classified within the definitions of the internet of things [1]. According to

N. K. Gupta (✉)
Department of Computer Science and Engineering, Jaypee University Anoopshahr, Anoopshahr, India
e-mail: nirmal.gupta@mail.jaypeeu.ac.in

M. K. Rohil
Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, India
e-mail: rohil@pilani.bits-pilani.ac.in

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_21

statistics, an average of 40,000 search queries occur every second; in other words, it can be said that over 3.5 billion searches are processed per day by Google search [2]. At the same time, various monitoring and sensing equipment are also generating data continuously. There is also a large amount of data in various fields such as scientific computing, healthcare, finance and retail. This phenomenon has aroused widespread concern.

The generation of this big data makes data analysis and application more complicated and difficult to manage. According to statistics, the amount of data generated globally over the past 3 years is more than the previous 400 years of data, which include documents, pictures, videos, web pages, e-mail, microblogging, and other types, which include mostly unstructured data as compared to structured data [3]. Gartner had forecasted that 4.9 billion connected objects to be in use by 2015, up 30% from 2014 and will reach 25 billion by 2020 [4]. At present, big data has become another information industry growth point in the field of information technology after cloud computing.

Like other current information systems, big data also involves security risks in the process of storage, processing, transmission and similarly it needs the data and privacy to be protected. The services like storage and management of data are provided in big data and cloud computing by the service providers themselves. But, the problem with big data as compared to cloud computing is that in case of cloud computing the user still has some control over their data to some extent; for example, through the use of cryptographic methods and through the other trusted computational methods. However, in the context of big data the businesses like Facebook not only produce the data but also provide the services like storage and management of data themselves. Therefore, it is extremely difficult to restrict the use of user information by means of technology and to protect the privacy of users' data [5].

At present, many organizations are aware of the big data security issues, and actively take action to focus on big data security issues. This paper focuses on the security challenges brought by the current big data technology and elaborates on the key technologies used for big data security and privacy protection. It should be pointed out that while introducing new security issues and challenges, big data also brings new opportunities in the field of information security. That is, big data-based technologies for information security can be used in turn for big data security and privacy protection.

2 Big Data Research Overview

2.1 Big Data Sources and Characteristics

Big data may have various sources from where it may be generated. Based upon the source of generation of big data, it can be divided into following categories [6]:

1. From the people: all kinds of data generated by people in the process of performing activities over the internet. The generated data can be in the form of text, images, videos or of any other type.
2. From the machines: This includes the data which is generated by different computers and information processing units which may be in the form of files, databases, multimedia, and so on, and also includes automatically generated information such as logs.
3. From the devices: data collected by various types of digital devices, such as the digital signals continuously generated by the camera; the different data related to human beings generated through various medical devices; the large amount of data which is generated by the astronomical telescopes.

2.2 Big Data Analysis Goal

At present, big data analysis is applied to various diversified areas such as science, medicine and commerce. Overall, the goals of big data analytics fall into the following categories:

(1) *Gain knowledge through extensive analysis*

People have a long history of data analysis. There may be various reasons. The first and most important reason for analyzing the data is to get knowledge from it. Since there is a large amount of unprocessed real sample information, it can effectively abandon individual differences and help people through the mining, and more accurately grasp the common purpose behind the things. Depending on the knowledge they have discovered, one can predict more accurately the nature or social phenomena that will occur. Typical examples include the ability to retrieve information about the flu through Google's search using data mining [7]; predicting stock quotes based on Twitter information [8]; and so on.

(2) *Grasp individual laws through long-term analysis*

Individual activities have distinct personal characteristics while satisfying certain common characteristics. Through long-term multi-dimensional data accumulation and gaining information through that data, various companies get the insights of users' behavior and this may help them to accurately describe the individual user's profiles. In this way, it helps them to provide more accurate products and services according to users' individual needs. It also helps companies to accurately provide recommendations related to advertisements.

For example, Google analyzes users' habits and hobbies through its big data products, helping advertisers to evaluate the efficiency of their advertising campaigns, and it is estimated that there may be hundreds of billions of dollars in the market in the future [9].

(3) *Control epidemic through analysis*

Many times the timely information obtained through the analysis of big data can provide more valuable information regarding spread of epidemics, than by disease-prevention centers. For example, during the 2009 flu pandemic, big data analysis was performed by Google to get the timely information. Generally, patients do not go to the doctor immediately just after getting infected, but their search and discussion trends can be analyzed to get information about most influenced geographical areas.

2.3 Big Data Technology Framework

Big data processing involves data collection, management, analysis and display. Figure 1 is a schematic view of the relevant technology, including four stages.

1. Data acquisition and preparation

The data sources of big data are diversified, including all kinds of structured, unstructured and semi-structured data such as databases, texts, pictures, videos and web pages. Therefore, the first step in big data processing is to gather data from the data source and pre-process it to provide a consistent, high-quality data set for subsequent processes.

Since there exist various sources of big data, therefore, there may be different models for its description and these may even contradict. Therefore, it becomes important to clean the data during the data integration process so that similar, repetitive or inconsistent data can be removed. In the literature, data cleaning and

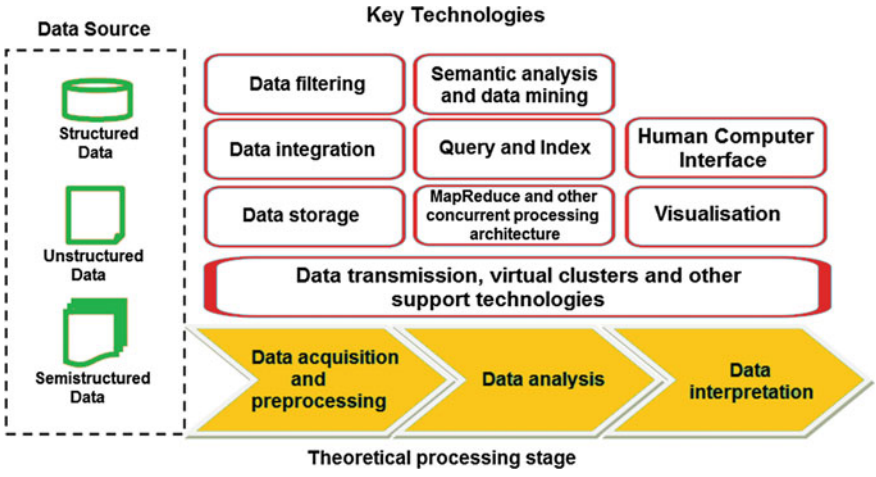

Fig. 1 Big data technology architecture

integration technology is aimed at the characteristics of big data, and proposes the cleaning of unstructured or semi-structured data and the integration of very large-scale data [9].

Data storage and big data applications are closely related. Real-time applications using big data require strong computational ability so that large amount of data could be processed in less amount of time. Therefore, big data processing system for real-time systems should quickly respond to the requests generated. This requires robust computing power for big data. Stream processing mode is more suitable for such applications. Most of the other applications need storage for subsequent deeper data analysis process. This may increase the storage cost. Usually big data uses distributed architecture to increase data throughput and reduce storage cost. Representative studies include file system GFS, HDFS and Haystack; NoSQL database Mongodb, CouchDB, HBase, Redis, Neo4j and so on [10].

2. Data analytics

Traditional data analysis may not work for big data as it was devised for structural data sources, but big data mostly consists of semi-structured or unstructured data. This presents a big challenge to big data analysis, but it is also the core process of big data applications. Based on the different levels, it can be broadly divided into three categories: computing architecture, queries and indexes, and data analysis and processing.

In terms of computing architecture, MapReduce [11] is a widely used big data set computing model and framework. In order to adapt to some analysis requirements that require high task completion time, work in [12] optimizes its performance. In [13], a data flow analysis solution based on MapReduce architecture, MARISSA, was proposed to support real-time analysis task. Dede et al. [14] proposed Mastiff, a big data analysis program based on time. Chandramouli et al. [15] proposed a TiMR framework based on MapReduce to deal with the real-time streaming for applications with high real-time demand such as advertisement push.

In query and indexing, traditional relational database query and indexing techniques are limited due to the large amount of unstructured or semi-structured data contained in big data, and NoSQL database technology received more attention. For example, Chandramouli et al. [16] proposed a hybrid data access architecture, HyDB, as well as a method of concurrent data query and optimization. Wang et al. [17] optimizes the query of key-value type database.

In data analysis and processing, the main technologies involved include semantic analysis and data mining. Owing to the diversification of data in big data environment, it is difficult to unify the terms to mine information when semantic analysis of data is concerned. In [18], for the big data environment, a high-efficiency terminology standardization method for solving the term variation problem is proposed. Keteta et al. [19] studied the heterogeneity of semantic ontology in semantic analysis. Traditional data mining technology is mainly aimed at structured data, so it is very important to study unstructured or semi-structured

data mining technology. Kang et al. [20] proposed a mining technique for image files, and Kang et al. [21] proposed a large-scale TEXT file retrieval and mining technology.

3. Data interpretation

The purpose of data interpretation is to represent data analysis results in a way which can serve the user's purpose. The major technologies which make it possible are visualization and human–computer interaction. There have been some visualization studies for large-scale data [22], which solve the display problem of large-scale data through data projection, dimension degradation or display wall. As human visual sensitivity limits the effectiveness of larger screens, a human-centric human–computer interaction design will also be an important technique to address the display of big data analytics results.

4. Other support technologies (data transmission and virtual cluster)

Although big data applications emphasize data-centric computing and push calculations to data execution, data transmission is still essential throughout the process, such as the transmission of some scientific observations from observation points to data centers. In [23] the authors study efficient transmission architectures and protocols for big data features.

In addition, because virtual clusters have the advantages of low cost, flexible construction and easy management, people can choose more convenient virtual clusters to complete the various processing tasks during big data analysis. Therefore, virtual machine cluster optimization research for big data applications is needed [24].

3 Big Data Security Challenges

Big data provides a great technology which has its significance in various fields and the security requirements in these fields are also changing. During the various activities performed over data during its collection, refinement and mining, many security threats are also associated. During this process the data may be destroyed, leaked, tampered, which can put the user privacy or corporate secrets to be compromised. In general, the security related to big data has the following characteristics and challenges.

3.1 Technical Challenge

The technical challenges of big data applications are mainly in the following three aspects:

First, as we all know, the application of big data is based on the possession of massive data. This involves the challenges of data storage technology and the technical challenges for data processing and analysis, including the data processing capabilities of computer hardware and supercomputer algorithm technology. The report learned from interviews with technical experts indicates that they are not optimistic about the recent overcoming of the technical challenges described above. Second, in the educational application of big data, data collection and problem-solving analysis are the core links, and application developers have to face the challenges of data acquisition technology and problem-solving analysis technology. The third aspect is the data compatibility challenge, the inconsistency of data encoding and format in different data storage systems, resulting in data sharing difficulties between different systems. The main reason for this problem is the lack of unified planning for the construction and purchase of various systems which results in inability to form unified data platform.

3.2 Mobile Data Security Challenge

In the world of today, various new applications related to social media have been emerged; besides this it is not uncommon for e-commerce portals and mobile applications to generate enormous amount of data. Analyzing such large amount of data in addition to data generated through internet of things network is a big challenge for companies. It also put responsibilities over the companies as the demand for data security capabilities on enterprises also increases greatly. In addition, the increase in data secrecy results in more sensitive analysis of data which is being transferred between mobile devices. Personal information may also be compromised because of the malware which may track user location or steals the confidential information. It requires increasing the safety related to user's privacy and data. Since the mobile devices have grown rapidly in recent years, it has created a challenge for big data security that how the samples of malware could be tracked which may enable analyzing these malware samples and finding the relationship between them.

3.3 Easy Target for Attackers

The various resources and personalized services are being provided through the network access in a flexible manner. In such a large networked society, the voluminous data attract the hackers because of the potential value associated with it. Such big data is large and has interrelated information stored, and due to this an attacker finds it easy to get more information by successfully attacking once, which further reduces the attacking cost and increases their profit value. A good example of such situation is when hackers use information in big data and take control of

important systems to take advantage from it for the time being. Hackers can also deploy AI technologies to take full control of the system by slowly penetrating it over time. Such strategy may work, as it would perform minor changes and such changes would appear natural. Thus, its detection could be avoided.

3.4 User Privacy Protection Challenge

Worldwide people use different kinds of web services and while using such services they need to access the network which stores their credentials like name, username, password, contact information, address and so on. It also includes their other personal information, behavior, habit, political or religious inclination and location; this has increased the risk of user privacy and data disclosure. Through the use of various data mining approaches such data can be easily recorded. If these data are found to be useful, then companies in the relevant areas of customer's need and habit can target to the specific users to achieve greater benefits. The traditional approach of privacy protection is based on anonymity of user data which is publicly available. But in reality, the user privacy protection cannot be achieved only through data protection through anonymity. There exist many other requirements and characteristics for user privacy. But the problem is that most of the existing privacy protection models and algorithms are only for traditional relational data, and cannot be directly applied to big data applications.

3.5 Safe Storage of Massive Data Challenge

The quantity of big data has increased tremendously, which includes structured and unstructured data. It has increased to such a level that the previous storage systems are not capable to meet the needs of big data applications. The current disk technology has some limitations and this limit is around 4 terabytes per disk. So for such a massive data of exabyte levels, this limitation creates a problem, because just to store 1 exabyte requires around 25,000 disks. It will really create the hardware issue of how to attach such large number of disks to a single computer, even if the computer is capable of processing that much of data.

To satisfy the demands of data storage of such large amount, various new storage technologies have been developed. Storage technologies such as direct attached storage (DAS), network attached storage (NAS) and SAN (storage area network) are being used to solve the data storage-related issues. Another technology called NoSQL storage technology is also used to capture, manage and process big data. Using NoSQL data storage can be extended and performance also can be improved, but still some issues exist. These issues include access control and

privacy control issues, issues related to technical vulnerabilities, security issues for authorization and verification, data management and confidentiality issues, and so on.

3.6 Challenge of Analysis

With the increase of storage capabilities, prior selection of data became insignificant. This can be seen as a real chance, allowing to keep focus on the potential future uses and which are not always fully defined at the time of their acquisition. In particular, many issues that were considered nonexistent later become accessible before the use of their potential of significant advantage (competitive advantage for example). It should be kept in mind, however, that more data is not always better data. It depends on whether or not they are heterogeneous, and whether they are representative of what is being sought. In addition, as the number of parameters increases, the number of erroneous correlations also increases. The analysis part will have to take into account these essential aspects.

Heterogeneous (structured, unstructured) data or incomplete or uncertain data for which specific treatments are needed will also be stored. Moreover, in this regard specific treatments are already required for the more standard data, even if some old methods remain effective for the volume of the existing data, which may cause theoretical and practical difficulties, unknown in advance. Thus, the simple statistical tests [25] become inoperative for large sample sizes. We can also mention the difficulty of multi-dimensional analysis on large sets of data, which arises during their interpretation. The visualization is considered until now as an extremely powerful tool, but it risks becoming inoperative by simple graphic saturation effect. In addition, real-time analysis of continuous flow of data from different sources also poses specific challenges. All these problems involve the development of new statistics for big data, for example, requiring a review of basic calculations such as statistical tests and correlations [26]. Of course, these technical analysis tools cannot be isolated from the computer tools and techniques dedicated to big data, for example, NoSQL, Hadoop, MapReduce or Spark.

3.7 Big Data Security Trust Challenge

Although big data has provided various opportunities to its users, but still it lacks the complete trust of the people using it. The visibility of social profiles generated by users varies across different types of networks and these are crawled by the search engines and therefore they become visible by the other users whether they have account or not. Therefore, here a trust issue arises from the user that how safe their data privacy is. This requires trust measures to be integrated with big data. These trust measures should not be treated as a static measure. That means, as the

data evolves the trust measures should also be updated accordingly. Yin and Tan [27] in their research have put the fact that semi-supervised learning methods that start with ground truth data are able to provide higher accuracy and trust on the source data. Another fact is that the different people have different personal opinions regarding various factors affecting their life and when there exist differences with statistics it leads to the market doubts about statistics.

4 Key Technologies for Big Data Security and Privacy Protection

At present, it is urgent to carry out research on key technologies of big data security in view of security challenges such as user privacy protection, data content credibility and access control faced by the big data. This section introduces some related key research areas for this.

4.1 Data Anonymity Protection Technology

For structured or unstructured data in big data, the core key technologies and basic means for data protection to achieve its privacy protection are still in the stage of development and improvement. Take the typical k anonymity scheme as an example. The early schemes [28] and their optimization schemes [29] group quasi-identifiers by data processing such as tuple generalization and suppression. The quasi-identifiers in each packet are the same and contain at least k tuples, so each tuple is at least indistinguishable from $k - 1$ other tuples. Since the k-anonymous model is for all attribute collections, it is not defined for a specific attribute, and it is prone to insufficient anonymity of a certain attribute. If the value of a sensitive attribute in an equivalence class is the same, the attacker can effectively determine the value of the attribute. This research is for static, one-time release. In reality, data publishing often faces scenarios in which data is continuously and repeatedly released. It is necessary to prevent an attacker from analyzing the data associations that are published multiple times, and destroying the original anonymity of the data [30].

In big data scenarios, data anonymity protection is more complicated: an attacker can get data from multiple sources, not just the same source. For example, in the Netflix application, people [31] found that an attacker could identify the target's Netflix account by comparing the data to the publicly available imdb. According to this, the user's political inclinations and religious beliefs are obtained (obtained through the user's viewing history and comments and scoring analysis of certain movies). Such issues are subject to further research.

4.2 Social Network Anonymity Protection Technology

The data generated by social networks is one of the important sources of big data, and it contains a large amount of user privacy data. Because social networks have the characteristics of graph structure, their anonymous protection technology is very different from structured data.

The social networks also require anonymity protection. Here, some typical requirements are user anonymity and attribute anonymity. While using social networks the user identity and attribute information is also required to be hidden while publishing these. The related data of different users should not disclose their relationship and anonymity between users is required. It is also known as edge anonymity. Hide the relationship between users when publishing. The attacker tries to use the various attributes of the node (degrees, tags, some specific connection information, etc.) to re-identify the identity information.

The current side-anonymity schemes are mostly based on additions and deletions of edges. The method of randomly adding and deleting exchange edges can effectively implement edge anonymity. Among them, Ying et al. [32] keep the eigenvalues of the adjacency matrix and the corresponding second eigenvalues of the Laplacian matrix in the anonymity process. Zhang et al. [33] group according to the degree of the nodes, and select the nodes with the same degree. The problem with this type of method is that the randomly added randomness is too scattered and sparse, and there is a problem of insufficient protection of the anonymous side.

4.3 Data Watermarking

Digital watermarking refers to a method in which identification information is embedded in a data carrier in an imperceptible manner without affecting its use, and is more commonly found in multimedia data copyright protection. There are also some watermarking schemes for databases and text files.

The method of adding watermarks in databases and documents is very different from the multimedia carrier, which is determined by the characteristics of data disorder and dynamics. The basic method is that there can be redundant information in the data or it can bear certain precision errors. For example, Agrawal et al. [34] have an error tolerance range based on numerical data in the database, embedding a small amount of watermark information into the least significant bits randomly selected from these data. Sion et al. [35] proposed a scheme based on statistical features of data sets, embedding one-bit watermark information in a set of attribute data to prevent attackers from destroying the watermark. In addition, by embedding database fingerprint information in the watermark [36], the owner of the information and the object being distributed can be identified, which is beneficial for tracking the leak in a distributed environment. Watermarking based on text content [37] depends on modifying the content of the document, such as adding spaces,

modifying punctuation, natural language-based watermarking [38] and so on, through the understanding of semantics to achieve changes, such as word substitution or sentence changes.

4.4 Data Traceability Technology

Owing to the diversified sources of data, it is necessary to record the source of the data and its distribution to provide support for later mining and decision-making. Data provenance technique has been extensively studied in the database field long before the big data concept emerged. The basic purpose is to help people determine the source of each data in the data warehouse, for example, which data items in tables are computed, so that it is convenient to check the correctness of the results at a very small cost. The basic method of data tracing is notation, such as marking the data in the data warehouse in [39] to record the query and propagation history of the data in the data warehouse. Data traceability techniques can also be used for traceability and recovery of files. For example, the work in [40] created a prototype system of data origin storage systems by extending the Linux kernel and file system, which can automatically collect origin data.

Further data traceability technologies can play an important role in the field of information security. However, data traceability technology also faces the following challenges in the protection of big data security and privacy:

1. *The balance between data traceability and privacy protection.* Using traceability to provide big data security protection requires first to obtain big data source using analysis of big data. Then the next step becomes to define the security policy and provide the required security mechanism. Often, the source of big data is privacy-sensitive and users are not interested that data to be accessed by the analysts also. Therefore, the problem is how to balance these two requirements simultaneously so that data traceability and privacy protection of the data both can be achieved.
2. *Security protection of data security technology itself.* The data tracing techniques currently employed are unable to handle the security issues correctly. The problem is that how to determine whether the tag associated with the data is itself correct or not. The other problem is that the tag information itself may not be securely bound with the data content and there may be other similar issues. Also in case of big data, since it is implemented on such a large scale, high speed and diverse characteristics such problems become more important.

5 Conclusion

The arrival of the era of big data has opened great opportunities. Big data not only has impact on everyone's social and economic behavior but also has influenced their way of living and thinking. Although big data is an important solution to various problems, it has also brought new security issues into existence. From the perspective of privacy protection, trust and access control of big data, this paper analyzes various security features and problems in the big data environment, namely, mobile data security, attack targets, user privacy protection challenges and security, storage issues, data security evolution, trust security issues and so on, and also discusses the preventive solutions for them. However, generally speaking, the current research on the protection of big data security and privacy is not sufficient. Only through the combination of technical means and relevant policies and regulations combined can better solve the big data security and privacy protection issues.

References

1. Xia, F., Yang, L.T., Wang, L., Vinel, A.: Internet of things. Int. J. Commun. Syst. **25**(9), 1101–1102 (2012)
2. Google search statistics. http://www.internetlivestats.com/google-search-statistics/
3. Lee, I.: Big data: dimensions, evolution, impacts, and challenges. Bus. Horiz. **60**(3), 293–303 (2017)
4. Nguyen, B., Simkin, L.: The Internet of Things (IoT) and marketing: the state of play, future trends and the implications for marketing. J. Mark. Manage. **33**(1–2), 1–6 (2017)
5. Boyd, D., Crawford, K.: Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. **15**(5), 662–679 (2012)
6. Guo-Jie, L., Xue-Qi, C.: Research status and scientific thinking of big data. Bull. Chin. Acad. Sci. **27**(6), 647–657 (2012)
7. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
8. Arias, M., Arratia, A., Xuriguera, R.: Forecasting with twitter data. ACM Trans. Intell. Syst. Technol. **5**(1), 8 (2013)
9. Arasu, A., Chaudhuri, S., Chen, Z., Ganjam, K.: Experiences with using data cleaning technology for bing services. IEEE Data Eng. Bull. **35**(2), 14–23 (2012)
10. Sarma, A.D., Dong, X.L., Halevy, A.: Data integration with dependent sources. In: Proceedings of the 14th International Conference on Extending Database Technology, ACM, pp. 401–412 (2011)
11. Elomari, A., Maizate, A., Hassouni, L.: Data storage in big data context: a survey. In: International Conference on Systems of Collaboration (SysCo), pp. 1–4. IEEE (2016)
12. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
13. Verma, A., Cherkasova, L., Kumar, V.S., Campbell, R.H.: Deadline-based workload management for mapreduce environments: pieces of the performance puzzle. In: Network Operations and Management Symposium (NOMS), pp. 900–905, IEEE (2012)

14. Dede, E., Fadika, Z., Hartog, J., Govindaraju, M., Ramakrishnan, L., Gunter, D., Canon, R.: Marissa: Mapreduce implementation for streaming science applications. In: IEEE 8th International Conference on E-Science (e-Science), 2012, pp. 1–8, IEEE (2012)
15. Guo, S., Xiong, J., Wang, W., Lee, R.: Mastiff: a mapreduce-based system for time-based big data analytics. In: IEEE International Conference on Cluster Computing (CLUSTER), 2012, pp. 72–80, IEEE (2012)
16. Chandramouli, B., Goldstein, J., Duan, S.: Temporal analytics on big data for web advertising. In: IEEE 28th International Conference on Data Engineering, pp. 90–101. IEEE (2012)
17. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf. Sci. **275**, 314–347 (2014)
18. Wang, Y., Lu, W., Wei, B.: Transactional multi-row access guarantee in the key-value store. In: IEEE International Conference on Cluster Computing (CLUSTER), 2012, pp. 572–575. IEEE (2012)
19. Hwang, M., Jeong, D.H., Jung, H., Sung, W.K., Shin, J., Kim, P.: A term normalization method for better performance of terminology construction. In: International Conference on Artificial Intelligence and Soft Computing, pp. 682–690. Springer, Berlin (2012)
20. Ketata, I., Mokadem, R., Morvan, F.: Biomedical resource discovery considering semantic heterogeneity in data grid environments. In Integrated Computing Technology, pp. 12–24. Springer, Berlin (2011)
21. Kang, U., Chau, D.H., Faloutsos, C.: Pegasus: mining billion-scale graphs in the cloud. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5341–5344. IEEE (2012)
22. Kola, A., More, H., Soderman, S., Gubanov, M.: Generating Unified Famous Objects (UFOs) from the classified object tables. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 4771–4773. IEEE (2017)
23. Tang, J., Liu, J., Zhang, M., Mei, Q.: Visualizing large-scale and high-dimensional data. In: Proceedings of the 25th International Conference on World Wide Web, pp. 287–297. International World Wide Web Conferences Steering Committee (2016)
24. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of "big data" on cloud computing: review and open research issues. Inf. Syst. **47**, 98–115 (2015)
25. Meyer-Schönberger, V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan, England (2013)
26. Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manage. **35**(2), 137–144 (2015)
27. Yin, X., Tan, W.: Semi-supervised truth discovery. In: Proceedings of the 20th International Conference on World Wide Web, pp. 217–226. ACM (2011)
28. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Syst. **10**(5), 557–570 (2002)
29. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM (2005)
30. Xiao, X., Taom, Y.: M-invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 689–700. ACM (2007)
31. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy, 2008. SP 2008, pp. 111–125. IEEE (2008)
32. Ying, X., Wu, X.: Randomizing social networks: a spectrum preserving approach. In: Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 739–750. Society for Industrial and Applied Mathematics (2008)
33. Zhang, L., Zhang, W.: Edge anonymity in social network graphs. In: International Conference on Computational Science and Engineering, 2009. CSE'09, vol. 4, pp. 1–8. IEEE (2009)
34. Agrawal, R., Haas, P.J., Kiernan, J.: Watermarking relational data: framework, algorithms and analysis. VLDB J. Int. J. Very Large Data Bases **12**(2), 157–169 (2013)

35. Sion, R., Atallah, M., Prabhakar, S.: On watermarking numeric sets. In: International Workshop on Digital Watermarking, pp. 130–146. Springer, Berlin (2002)
36. Guo, F., Wang, J., Li, D.: Fingerprinting relational databases. In: Proceedings of the 2006 ACM Symposium on Applied Computing, pp. 487–492. ACM (2006)
37. Pease, A., Niles, I., Li, J.: The suggested upper merged ontology: a large ontology for the semantic web and its applications. In: Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, vol. 28, pp. 7–10 (2002)
38. Atallah, M.J., Raskin, V., Hempelmann, C.F., Karahan, M., Sion, R., Topkara, U., Triezenberg, K.E.: Natural language watermarking and tamperproofing. In: International Workshop on Information Hiding, pp. 196–212. Springer, Berlin (2002)
39. Cui, Y., Widom, J., Wiener, J.L.: Tracing the lineage of view data in a warehousing environment. ACM Trans. Database Syst. **25**(2), 179–227 (2000)
40. Muniswamy-Reddy, K.K., Holland, D.A., Braun, U., Seltzer, M.I.: Provenance-aware storage systems. In: USENIX Annual Technical Conference, General Track, pp. 43–56 (2006)