Privacy and Security in Big Data, Categories, Issues, and Proposed Solutions

Younus AbdulKareem Mohammed
Iraq University College – Iraq -Basra
younusyounus50@gmail.com

Abstract

In recent years, one of the most contentious issues has been "big data." Unstructured and semi-structured data may be found in many locations, including web servers, mobile devices, and social media sites, for many organizations and enterprises, the security and privacy of large data are becoming more critical, and the increased usage of big data puts data security and privacy at odds. Large-scale data sharing raises new privacy and security issues. Big data doesn't work with traditional methods or techniques. Analyzing large amounts of data encourages the collection and long-term storage of more complete and durable data. By combining private data with other private data, it is possible to reveal the personal information of its customers more easily. The sharing of massive data opens the door to new problems. It necessitates the use of high-tech methods and tools because they create so many data security issues. As part of our review of Big Data privacy and security the most recent methods, mechanisms, and solutions for protecting data-intensive systems were discussed. we looked at the most important terms to define and classify, in this paper, big data security and privacy researchers will benefit from this review because it identifies key trends and general terms, as well as current privacy and security concerns. And we also highlighted common solutions to these problems included.

Keywords: Privacy, Security, Big Data.

الخصوصية والأمان في البيانات الضخمة ،الفئات ،المشكلات ،الحلول المقترحة

يونس عبد الكريم محمد
كلية العراق الجامعة - العراق - البصرة

الخلاصة

في السنوات الأخيرة ، كانت "البيانات الضخمة" من أكثر القضايا إثارة للجدل، يمكن العثور على البيانات غير المهيكلة وشبه المهيكلة في العديد من المواقع ، بما في ذلك خوادم الويب والأجهزة المحمولة ومواقع الوسائط الاجتماعية ، وبالنسبة للعديد من المؤسسات والشركات ، أصبح أمان وخصوصية البيانات الكبيرة أكثر أهمية ، وزيادة استخدام البيانات الضخمة يضع أمن البيانات والخصوصية على خلاف. تثير مشاركة البيانات على نطاق واسع قضايا خصوصية وأمان جديدة ، لا تعمل البيانات الضخمة مع الأساليب أو التقنيات التقليدية. يشجع تحليل كميات كبيرة من البيانات على جمع وتخزين بيانات أكثر اكتمالاً واستمرارية على المدى الطويل. من خلال الجمع بين البيانات الخاصة والبيانات الخاصة الأخرى ، من الممكن الكشف عن المعلومات الشخصية لعملائها بسهولة أكبر، تفتح مشاركة البيانات الضخمة الباب أمام مشاكل جديدة. يستلزم استخدام أساليب وأدوات عالية التقنية لأنها تخلق الكثير من مشكلات أمان البيانات،كجزء من مراجعتنا لخصوصية وأمن البيانات الضخمة ، تمت مناقشة أحدث الأساليب والآليات والحلول لحماية الأنظمة كثيفة البيانات. نظرنا في أهم المصطلحات التي يجب تعريفها وتصنيفها ، في هذه الورقة ، سيستفيد باحثو أمن وخصوصية البيانات الضخمة من هذه المراجعة لأنها تحدد الاتجاهات الرئيسية والمصطلحات العامة ، فضلاً عن مخاوف الخصوصية والأمان الحالية. وقمنا أيضًا بتسليط الضوء على الحلول المشتركة لهذه المشاكل.

الكلمات المفتاحية: الخصوصية ، الأمان ، البيانات الضخمة.

1. Introduction

Streaming cloud technology has enabled the growth of Big Data, and traditional security mechanisms designed for small-scale, static data on firewalled and semi-isolated networks are inadequate. Examples of excessive outliers would be generated by analytics used to detect anomalies. As with provenance, it's unclear how it can be retrofitted into the current cloud infrastructure. Security and privacy solutions that provide lightning-fast response times are especially important when dealing with streaming data (Alliance for Cloud Security, 2013). Because so few people are familiar with the processes involved in implementing security and privacy features, most current technologies are deficient. However, in light of current data protection and privacy concerns, the most security and privacy programs are not always feasible(Ghosh and Nath, 2016). Massive amounts of personal data are mined using artificial intelligence and machine learning in Big Data analysis to find patterns that can be used to inform decisions at the individual level. This raises privacy issues, as well as issues relating to due process, discrimination, and consumer protection. Following Paterson and McDonagh (2018, for instance), It's true that Big Data is a major concern for governments and businesses around the world, but so far there have only been a few successful Big Data implementations. As Big Data solutions are implemented, several challenges will need to be addressed, including the security of personal information and accurate estimates of financial resources required for large-scale infrastructure investment It's important to note that Paterson and McDonagh (2018) Also, the issue of security and privacy. Health care, transportation, and even the entertainment industry are among the many fields where Big Data could make a significant impact, but they are also subject to stringent security and privacy regulations. To outsource data is to compromise your privacy and safety. Data and computing migration away from the internal IT infrastructure increases user threats, such as data loss, and data breach and increases the chances of corporate asset theft (Ardagna et al., 2016).

2. Privacy and security are two important considerations:

Securing the privacy of big data is critical before we can use it for more general purposes (Yu, 2016). We are compelled to enhance our current data privacy practices as new and innovative technologies emerge regularly. Users become aware that they are adding data to the system when they utilize online services such as email, social networking, and news feeds. As if that weren't awful enough, data may be sold for analysis to a third party. The implementation of privacy standards, new structures, and safeguards is projected to be comprehensive.

To secure Big Data, certain rules and procedures should be in place throughout its lifespan. Each stage of the generic data life cycle, from collecting through filtering, processing, and analysis, is a positive step. After then, the user is given a graphical depiction of the data (Demchenko et al., 2014). Allowing user hazards such as data loss and breaches to enter the company's IT infrastructure. In multi-source, dispersed environments, big data analytics has extra challenges. To accept big data, a business must also consider the demands of its owners (Ardagna et al., 2016). When it comes to analyzing and aggregating enormous volumes of data, personal and private information, intellectual property, corporate secrets, and trade secrets are among the most

sensitive and valuable assets that organizations have at their disposal. Businesses are increasingly deriving value from information regarding their stakeholders' relationships with them. Determining how to store, process, and analyze massive amounts of "big data" has become critical for all parties engaged in the chain of events. And to get the most value out of big data, it must be processed and assessed quickly, with the findings influencing business decisions. People, processes, and technology can only help organizations reach their maximum potential(Kapil et al., 2018). The confidentiality, integrity, and validity of information systems must be safeguarded from external assault. As a result, a secure system is necessary to safeguard the integrity, identity, and privacy of people's data(Kapil et al., 2020). however, cyber security applications may face adversaries that actively adapt their methods. In contrast to other sectors that use data analytics tools, numerous firms now provide data analytics solutions to address this essential problem. The objective is to provide cyber security professionals with prioritized actionable insights from large data. Nonetheless, using data analytics for cyber security may be incorrect. In contrast to the majority of other application domains (Kantarcioglu & Ferrari, 2019).

3. Methodology of Research

The research method used was a kind of literature search of several online libraries, including springer, SCOPUS, and the IEEE Digital Library. Google Scholar, Research Gate, the rationale for choosing these libraries over others than others were that they contained a substantial amount of literature pertinent to our primary objective and thus facilitated a targeted search. To ensure a proper outcome, a research string was created the following:

"Privacy and Security in Big Data, Categories, Issues, Proposed Solutions"

The first section of the research string focuses on the security and privacy implications of Big Data technology. The first part of the string discusses the various categories of privacy and security, the next part discusses the various categories of privacy and security, as well as the current issues and security concerns surrounding the topic of Big Data analytics, and the last part discusses possible solutions that can address your specific privacy and security concerns. Within the realms of Big Data and Big Data Analytics

4. There are several types of privacy and security concerns:

Big Data Working Group divided the examples into their primary privacy and security dimensions and used Cloud Security Alliances to illustrate the key points. 4.1 Big Data Security Types: Given the large number of Hadoop installations, it's critical to discuss the primary technologies and Hadoop-based ones when talking about Big Data infrastructure security.

a. **Security for Hadoop:** Hadoop can be used to help a business achieve Big Data. As a result, numerous research has proposed various security methods for enhancing the system's security. This is probably the most arbitrary method because it requires multiple mechanisms, such as authentication and cryptography, to ensure its security. rent categories of privacy and security.

b. **b Availability:** Researchers have taken on the responsibility of managing access to abundance. In big data environments, hundreds of computers are used, and the data is not only readily accessible but also replicated across the cluster.

c. **Architecture:** To deal with new Big Data, a completely different approach is to rethink the structure or to make it more secure. It enables increased system security using multiple reading nodes. Finally, there is another way to address the issue of secure group communications in big data networks, which involves modifying the system's nodes and protocols.

d. **Authentication:** The true value of data can be determined to a large extent by running a Big Data process.

e. **Security of Communication:** there is a communication breakdown between various components of the Big Data ecosystem the security is regarded as communications between different parts of the Big Data infrastructure, and the majority of solutions do not address this issue (Moreno et al., 2016).

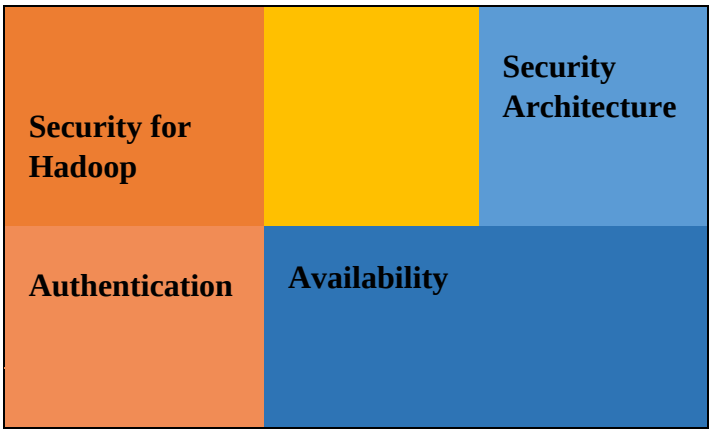

Figure (1) The Main topic of data security from ((Moreno et al., 2016)).

4.2 Data Privacy:

While most people are concerned about data privacy, any business that uses Big Data techniques should be particularly concerned. Typically, the data we obtain from a Big Data system contains a significant amount of personal information. the main categories of Data Privacy are:

a. **Encryption:** The most frequently used strategy in a Big Data system is data privacy. It has long been used to maintain the secrecy of information. While large data sets continue to pose this challenge, their characteristics preclude the application of traditional encryption techniques.

b. **Management of Permissions:** The access control principle is a long-standing practice in information security. the priority should be to safeguard the system against undesirable users. if Big Data is used.

c. **Confidentiality:** Historically, privacy has been considered a subset of confidentiality, but confidentiality has displaced privacy because of the enormous impact that Big Data technology has had on public perceptions. 4.2.4 Privacy-Preserving Queries While this is critical, the primary function of a Big Data system is to extract information from massive amounts of da

d. **Anonymization:** Anonymization is a highly effective method of protecting personal information. This includes any method of erasing or concealing personal information

from raw data. The problem is associated with ever-increasingly massive amounts of data, which results in ever-increasing amounts of data in ever-increasingly massive environments. However, while processing that data, we cannot overlook its privacy.

e. **Privacy in Social Networks:** Social networks pervade every aspect of our lives. Social networking sites have exploded in popularity, with nearly everyone on the Internet having a Facebook, Linked In, or Twitter account. They share a great deal of personal information but never consider who will use it once they have it. The combination of big data and creative thinking poses a serious threat to our privacy.

f. **Differential Privacy:** Differential privacy aims to maximize the value of data while minimizing the likelihood that users' data will be used to identify them (Moreno et al., 2016).

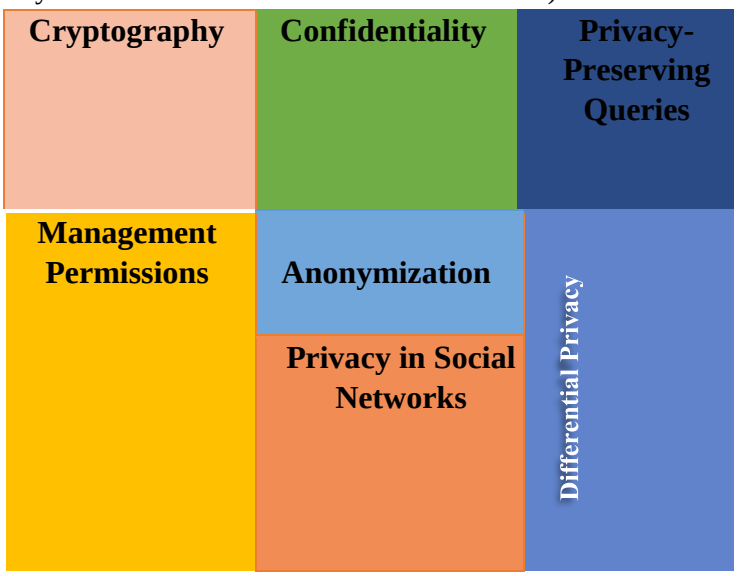

Figure (2) The Main topic of Data Privacy from (Moreno et al., 2016)

5. The current privacy and security issues on the topic of Big Data analytics:

According to previous studies, the most challenges to Big Data security and privacy have been identified as below:

(1) programming for distributed applications that incorporates security. (2) A database that is not relational (3) Secure data is always protected and transactions are secure. (4) Filtering for positive endpoints (5) real-time intrusion detection provides security. (6) data mining and analytics that are scalable and configurable. (7) Systemic data protection, or data security that is cryptographically enforced. (8) control on a very fine granular scale (9) auditing that is extremely thorough (10) primary sources of information (Cloud Security Alliance, 2013).

5.1 privacy Issues: many works of literature reviewed have identified several problems that arise with the advent of Big Data: most interesting issues are:

It promotes the acquisition and records its retention of more comprehensive and long-term data. Combining private data and merging it with other private data also assists in revealing consumer data. Most people are unaware of the amount of information they disclose when they use the Internet. According to economists and social scientists, these platforms foster a great deal of opportunism because of the resulting information asymmetries. It has

the potential to elevate an individual's status by exploiting existing advantages and vulnerabilities. Those with a higher level of education or wealth receive preferential treatment, which slows social mobility and advancement. Without adequate safeguards, it is possible for arbitrary and potentially discriminatory decision-based targeting and guilt-presumptive classifications to be made using correlated data. Since governments rely on administrative automation, citizens can be monitored and possibly harmed. Over time, self-determining thinking and self-determination can become discouraged. This "devaluing of humanity," in which agents' behavior is shaped involuntarily, is extremely detrimental to both society and individual welfare (Domingo-ferrer et al., 2015).

5.2 Data security issues:

the five most pressing big data security issues facing your organization, and how to address them the following are the obstacles we face: imaginative theory Insecure Computations: Cybercriminals extract and decrypt valuable information from confidential data sources using unknown computation methods. While external threats may steal your data, internal threats can also corrupt your results. To validate data as it is received and Filtering: Because big data must originate from multiple sources, it is critical to ensure its accuracy before integration. This should be done to ascertain which data is trustworthy and which is not. Additionally, it must protect user data from both malicious and non-malicious users. Granular specific authorization: Relational database stores extremely detailed and comprehensive data, whereas large-scale databases typically lack table- and cell-level control. Sensitive data from big data solutions are nearly impossible to obtain based on big data queries.

Data Storage and protection methods:

with so many data storage locations, authentication, authorization, and data encryption become significantly more difficult. Auto-tiering – The process by which fewer priorities data is moved to a less secure medium. If a solution provides real-time data encryption, it may be of limited utility if the encryption is compromised. Concerns about data loss may be raised by data mining and analysis. possible privacy concerns that may arise because of the use of data mining and analytics the dissemination of big data encompasses data mining and analytics, as well as a variety of issues such as privacy invasions, data theft, and blind market targeting. There are numerous instances of these types This could include AOL making public previously anonymized user search logs, which is terrifying (Patel & Prajapati, 2018).

6. The possible solutions that can address your specific privacy and security issues In Big Data and Big Data Analytics:

Because of the revolution of a multiplicity of connections, Big Data systems monitoring, and securing are more to become the responsibility of all stakeholders involved in systems (e.g., managers, security chiefs, auditors, end-users, and customers). As a result, it is advantageous to raise awareness of digital security practices and procedures among all parties and associated entities. It is necessary to incorporate security. it is everyone's responsibility to collaborate and ensure its security. here are some most common solutions for the privacy and security of big data.

6.1 The risk analysis for a variety of technologies

(Kim et al., 2013)suggests safeguarding the data with encryption rather than the data itself. If protecting a large amount of data is difficult, it cannot be used, and if it cannot be used, it is useless. additionally, big data security techniques make use of

attribute information. In essence, data owners or operators are responsible for specifying, selecting, and securing only critical attributes. In the context of Big Data, such critical data attributes receive this treatment:

- Contrast and contrast the characteristics, as well as their relationship to one another.
- Security mechanisms should be applied to those attributes that warrant protection,
- Choose security mechanisms to protect the relevant attributes following the owner's or organization's policies.

Big Data makes use of a variety of analytical solutions, including those from (Accenture, HP, EADS, CISCO, IBM, and Unisys). Those solutions provide increased flexibility and a slew of benefits, including enabling agile decision-making and rapid response, as well as assisting in the detection of attacks using real-time active and passive information (a low number of false positives) (Benjelloun & Lahcen, 2015).

6.2 Anonymization of Confidential or Personal Data

Anonymization of data in the cloud and distributed systems is a well-established technology. This strategy employs a variety of models and approaches, including the following: (Sub-tree data anonymization, closeness, m-invariance, k-anonymity, and l-diversity). Subtree techniques are based on two distinct strategies: Top-Down Specialization (TDS) and Bottom-Up Generalization (BUG). However, those methods are impractical. (Zhang et al., 2014) suggests combining the two BUG and TDS techniques (Blind User anonymization techniques, as a result, this approach compiles and applies the optimal strategy for the given set of parameters. This technique combines the necessary flexibility, productivity, performance, and scalability to efficiently and control large databases. It incorporates the MapReduce paradigm through recent modifications. When used properly, it enables distributed systems to run faster.

6.3 Data Cryptography

Numerous applications rely on data encryption to ensure both data security and the integrity of big data. Homomorphic cryptography enables computation on data that has been encrypted using conventional algorithms. according to this solution (Chen & Huang, 2013). suggested that an adapted platform or software environment deal with MapReduce computations in the case of Homomorphic Cryptography. to be certain of the performance of the cryptographic solutions in distributed environments (Liu et al., 2013). suggest a new approach for a key exchange called (CBHKE) (Cloud Background Hierarchical Key Exchange). It is a kind of secured solution used that is more rapid than its predecessor techniques (IKE and CCBKE). It is based on an iterative strategy to an (Authenticated Key Exchange) (AKE) through two phases (layer by layer).

6.4 Centralization and Security

The previous studies have favored the use of cloud computing devices over mobile ones. the objective is to leverage the cloud by utilizing normalized and standardized security mechanisms, as well as decentralized and synchronized systems. Without a doubt, the Cloud is constantly being improved and is constantly monitored for increased security. However, maintaining zero risks is difficult. Data security is a collaborative effort between humans. due to the Cloud's centralized storage of valuable data stay kind of secured. The objective is to hold all parties accountable for security management to increase the adoption of security best practices and to ensure compliance with applicable standards and laws. Users should be aware of potential threats, regulatory requirements, and

organizations' policies. (Benjelloun & Lahcen, 2015) .

6.5 Confidentiality and Monitoring of Data Access

We are seeing an increase in security threats because of the growth of distributed systems and the risk associated with data exchanges. To address these new security threats, the solution was to refine controls and segregate access levels by role, resulting in a plethora of control limits. There are numerous options for access and data security, such as single sign-on, credentials, federated identity management, and multi-factor authentication. To establish a protected exchange infrastructure for the exchange of protected data between agencies, three levels of authentication are available: public key infrastructure (PKI), and private key infrastructure (PKI) (certification authority, users, and machines). that is, to gain access to sensitive applications, a certificate in addition to a password is required. Federated Identity Management provides control and security enhancements for INDECT. This type of delegated domain-wide management is delegated to an identity provider (IdP). Security certificates and smart cards are the two primary tools in its security strategy. It is critical to safeguard data against prying eyes (Benjelloun & Lahcen, 2015).

6.6 preventative and threat detection

Constant surveillance is critical for detecting security incidents, threats, and other unknown behaviors as quickly as possible. To aid in the protection of Big Data surveillance, several strategies are available, including data loss prevention (DLP), information and event management (firewall detection/prevention), and security investigations. The methods used are data augmentation and contextualization, which make use of a variety of sources (to add context as a data attribute to the extracted data). It is equally critical to conduct regular audits and verifications of security policies and procedures (Benjelloun & Lahcen, 2015).

7. Conclusion:

when organizations integrate and share data, privacy and security concerns must be addressed. No aspect may be missed when it comes to safety and privacy. To fully use the potential advantages of big data, it is vital to consider many security and privacy concerns. At each stage of the process, the security and privacy of big data should be re-examined. When huge volumes of data are saved and recorded, privacy-conscious access control techniques must be implemented. On the other hand, The more data that is shared, the greater the risk of privacy and security challenges. This paper discussed several critical Big Data security and privacy issues based on the literature review. highlighting was done on the most serious security and privacy issues associated with big data and overviewed common solutions of data processing improvement used to make systems more reliable. It is easier to protect the data itself and its critical attributes in big data projects than it is to use a single big data technology for everything.

References:

(1) Ghosh, K., & Nath, A. (2016). Big Data: Security Issues and Challenges Big Data: Security Issues and Challenges. June.

(2) Kapil, G., Ishrat, Z., Kumar, R., Agrawal, A., & Khan, R. A. (2020). Managing Multimedia Big Data: Security and Privacy Perspective. Advances in Intelligent Systems and Computing, 1077(March), 1–12. https://doi.org/10.1007/978-981-15-0936-0_1N

(3) Kapil, G., Agrawal, A., & Khan, R. A. (2018). Big Data Security and Privacy Issues. Asian Journal of Computer Science and Technology, 7(2), 128–132.

https://doi.org/10.51983/ajcst-2018.7.2.1861
(4) Cloud Security Alliance. (2013). Expanded Top Ten Big Data Security and Privacy Challenges. Cloud Security Alliance, April 1–39.
(5) Paterson, M., & McDonagh, M. (2018). Data protection in an era of big data: The challenges posed by big personal data. Monash University Law Review, 44(1), 1–31.
(6) Paterson, M., & McDonagh, M. (2018). Data protection in an era of big data: The challenges posed by big personal data. Monash University Law Review, 44(1), 1–31. https://www.monash.edu/__data/assets/pdf_file/0009/1593630/Paterson-and-McDonagh.pdf
(7) Ardagna, C. A., Ceravolo, P., & Damiani, E. (2016). Big data analytics as-a-service: Issues and challenges. Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, 3638–3644. https://doi.org/10.1109/BigData.2016.7841029
(8) Yu, S. (2016). Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. IEEE Access, 4, 2751–2763. https://doi.org/10.1109/ACCESS.2016.2577036
(9) Domingo-ferrer, J., David, S., & Hajian, S. (2015). Privacy in a Digital, Networked World. January, 9–36. https://doi.org/10.1007/978-3-319-08470-1
(10) Demchenko, Y., Ngo, C., De Laat, C., Membrey, P., & Gordijenko, D. (2014). Big security for big data: Addressing security challenges for the big data infrastructure. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8425 LNCS, 76–94. https://doi.org/10.1007/978-3-319-06811-4_13
(11) Patel, H., & Prajapati, P. (2018). International Journal of Computer Sciences and Engineering Open Access. International Journal of Computer Sciences and Engineering, 6(10), 628–632.
(12) Moreno, J., Serrano, M. A., & Fernández-Medina, E. (2016). Main issues in Big Data security. Future Internet, 8(3). https://doi.org/10.3390/fi8030044
(13) Kim, S. H., Kim, N. U., & Chung, T. M. (2013). Attribute relationship evaluation methodology for big data security. 2013 International Conference on IT Convergence and Security, ICITCS 2013. https://doi.org/10.1109/ICITCS.2013.6717808
(14) Patil, H. K., & Seshadri, R. (2014). Big data security and privacy issues in healthcare. Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014, November, 762–765. https://doi.org/10.1109/BigData.Congress.2014.112
(15) Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., & Chen, J. (2014). A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. Journal of Computer and System Sciences, 80(5), 1008–1020. https://doi.org/10.1016/j.jcss.2014.02.007
(16) Chen, X., & Huang, Q. (2013). The data protection of mapreduce using homomorphic encryption. Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 419–421. https://doi.org/10.1109/ICSESS.2013.6615338
(17) Liu, C., Zhang, X., Liu, C., Yang, Y., Ranjan, R., Georgakopoulos, D., & Chen, J. (2013). An iterative hierarchical key exchange scheme for secure scheduling of big data applications in cloud computing. Proceedings - 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2013, 9–16. https://doi.org/10.1109/TrustCom.2013.65
(18) Kantarcioglu, M., & Ferrari, E. (2019). Research Challenges at the Intersection of Big Data, Security and Privacy. In Frontiers in Big Data (Vol. 2). https://doi.org/10.3389/fdata.2019.00001.