The ISC Int'l Journal of Information Security

November 2020, Volume 12, Number 3 (pp. 95–99)

http://www.isecure-journal.org

Selected Paper at the ICCMIT'20 in Athens, Greece

Knowledge Discovery in Cloud-Computing Environment and the Challenges**

Najah K. Almazmomi [1,*],

[1] *Department of Information System, College of Business, University of Jeddah, Jeddah, Saudi Arabia.*

ARTICLE INFO.

Keywords:
Knowledge Management,Knowledge Discovery, Cloud Computing, Big Data

Type: Research Article

doi: 10.22042/isecure.2021.273750.640

ABSTRACT

Today, in the area of tele-communication, social media, internet of things (IoT) and virtual world, enormous amounts of data are being generated which are extracted to discover knowledge. Knowledge discovery from data in the cloud-computing environment entails the extraction of new and necessary information from large and complex dataset. This study is qualitative and exploratory in nature. To review based on the recent literature, the articles published in the last five years (2014-2018) were searched. Different database were searched using the key words: "knowledge management" or "knowledge discovery" and "cloud computing". The literature review section is divided into three sub-section based on the findings. The first two sub-sections present the data security and data privacy concerns under two main techniques (Big data analytics and machine learning) used in knowledge discover; and the last sub-section presents various protocols proposed to address the related security and privacy concerns.

1 Introduction

Today, in the area of telecommunication, social media, internet of things (IoT) and virtual world, enormous amounts of data are being generated which are extracted to discover knowledge [1]. Knowledge discovery from data in the cloud-computing environment entails the extraction of new and necessary information from large and complex datasets. The information found through knowledge discovery can have various sources, such as data generated from social media, telecommunication logs, scientific experiments, internet of things (IoT), and surveys. The knowledge discovery is a serendipitous process, yet for discovering new knowledge, the right data and factors need to be present and aligned [2]. Knowledge discovery is a complex process and a subject of multidisciplinary domain that includes computer science, statistics and visualization [2]. For knowledge discovery, the experts apply various techniques to obtain information, such as programming, data mining, machine learning, visualisation and visual analysis, and human-computer interactions. In the context of cloud computing, the idea of knowledge discovery makes information as utilities that let the organisations to increase their functionalities and enhance their performance [3]. For that reason, knowledge discovery has both strategic and monetary importance for large business organisations, government organisations and research organisations

* Corresponding author.

**The ICCMIT'20 program committee effort is highly acknowledged for reviewing this paper.

Email address: Nalmazmomi@uj.edu.sa

ISSN: 2008-2045

[2]. Cloud computing involves simultaneous computing resources sharing by delivering parallel processing power and data storage over the Internet [1]. Similarly, for knowledge discovery, cloud-computing environment supports complex mining of large datasets with high scalability to gather information [1, 4]. Generally, the knowledge discovery process is applied in the domains of market analysis, business processes optimisation, economics, healthcare and public health, environmental study, bio-sciences, physics and various other sciences and research [4]. As most of the times, the cloud computing services are delivered by third party cloud vendors (such as Amazon) using their proprietary cloud software and data centres that are hosted outside the client organisations, the issues of service outage and data security and data privacy are very important to be concerned [5]. While the data security is about the measures and technology used to prevent unauthorized access, the data privacy is referred as authorized access, regarding who have it and who define it. Based on this concept, this paper provides an overview of the contemporary data security and privacy challenges related to different knowledge discovery techniques, and the protocols proposed to address these.

2 Methodology

This study is qualitative and exploratory in nature. To review based on the recent literature, the articles published in the last five years (2014-2018) were searched. Different database were searched using the key words: Knowledge management or Knowledge discover* and Cloud computing. Because of the time limitation, only the titles of the articles were screened, instead of going for abstract and full-text screening, to include the related articles for review. Subsequently, the selected articles were reviewed from the data security and privacy perspective. Following the thematic analysis method, the contents and findings of the articles were grouped based on the common techniques that are used for knowledge discover, and proposals that address the data security and privacy issues. For this reason, the literature review section is divided into three sub-section based on the findings. The first two sub-sections present the data security and data privacy concerns under two main techniques (Big data analytics and machine learning) used in knowledge discover; and the last sub-section presents various protocols proposed to address the related security and privacy concerns.

3 Lietrature Review

3.1 Big Data Analytics Technique

Today, newer technique such as Big data analytics does data mining for knowledge discovery in the cloud computing environments. This technique poses specific and significant security and privacy challenges [6]. For example, when GPS data from mobile devices and the data generated from Internet of Things technology (IoT) are integrated with Cloud Computing technologies for knowledge discovery, it becomes challenging to ensure that secured and encrypted communications are established during data transmission and relay among the devices [7]. According to the study by [8], cloud-computing technology is not a perfect solution yet, therefore, many companies are still reluctant to store their sensitive data outside their premises because of security and privacy concerns resulting from the data mining which can take place in the cloud-computing environments. To address this concern, it is important for the companies that security models and algorithms are developed. The security and privacy concerns in Big data, driven by the emerging concepts and applications of IoT, smart wearable and home devices, and smart cities, need to be addressed within the purview of steps in data collection, storing, sharing and accessibility [6]. According to European Union Commission report, privacy are the most important challenges in IoT design and research [9, 10]. For this reason, in the application of IoT technology, asking and getting users consent is significant issue to inform the users how they are affected by the devices or services. Parallelly, for the IoT security, the four important stakeholders, 1) the device manufacturer, 2) IoT cloud services and application provides, 3) the Government and Regulatory bodies, and, 4) the users, need to cooperate together [9]. According to Morsy [11], to address the security issues, the cloud computing models need to be devolved from the users perspective, cloud stakeholders perspective, and the cloud service delivery models perspective.

3.2 Machine Learning Technique

Machine learning is another technique used for knowledge discovery in the cloud-computing environment. This technique also uses data mining for knowledge discovery, however, it involves the use of algorithm that evolves automatically over time through gathering experience based on data. The scope of machine learning is now pervasive in every domain that involve digital or IT-enabled services. Still, because of security and privacy concerns, the full scope of machine learning is yet to be materialised and availed [12]. While many beneficial and personalised healthcare services are based on machine learning, which entails information harvesting and data intensive studies to extract meaningful information and obtain new knowledge, the associated privacy, data protection and data security issues are also becoming the concerns of utmost significance [13]. For this reason, although in the

nascent area, the cloud-computing organisations are continuously working to newer cloud-computing models to protect the privacy of individuals. As the third party cloud servers are generally not trusted enough, hence confidentiality and privacy issues are barriers to the full adoption of machine learning [14]. According to the study of Arpaci [15], privacy and security are important antecedents in cloud computing adoption by educational institutes for knowledge discovery and knowledge management. Thus, to take advantages of the unprecedented value of data mining and machine learning in the cloud-computing environment, where the medical records are used that contain personal health information, it is essential to develop and implement new policies to address the privacy issues carefully and systematically[16].

3.3 Protocols Proposed

As the new technologies such as cloud-based data mining are emerging, the privacy challenges are getting complicated and multi-dimensional, such as location privacy and search query privacy [17]. For example, in the use of Mobile Cloud Computing (MCC), location privacy is a key issue [18]. To address the privacy challenge related to data search query in public cloud, [19] used multi-keyword ranked search over encrypted cloud data (MRSE). Similarly, Pasupuleti, Ramalingam, and Buyya [20] used Efficient and Secure Privacy-Preserving approach (ESPPA) based on encryption technique to ensure data privacy in keyword search. On the other hand, to address the privacy-preserving for the users,[14] used multi-key fully homomorphic encryption (MK-FHE) method in deep learning process that is used for knowledge discovery. To address privacy concerns,[14] proposed the Privacy-preserving Outsourced Classification in Cloud Computing (POCC) model to address the privacy concerns. Similarly, according to the study of Fu, Shu, Wang, Liu and Lee [21], strong symmetric encryption algorithm is needed for ensuring the data security in the cloud-computing environment. Then again, according to [22], to ensure the safety and confidentiality of personal health records in the cloud-computing environment, data need to be protected from the external attackers and as well as the internal intruders. For this reason, they also suggested for encrypted data format in the cloud platforms. Liu, Au and Huang [23]have proposed fine-grained attribute-based two-factor authentication access control system and demonstrated its practicability in ensuring user privacy. In contrast, Yi, Rao and Bertino [24] proposed the distributed ElGamal cryptosystem that is practically useful for achieving privacy in both data storage and data transaction. Similarly, Yu, Au, Ateniese, Huang and Susilo [25] have proposed for Remote Data Integrity Checking (RDIC) technique to ensure zero knowledge privacy. The RDIC technique is found to be practical for real-world application in cloud-computing environment. The factors of security and privacy in the cloud computing environments are related to both hardware and software used in the cloud service and delivery models[26]. Although experts are proposing and developing various cloud architecture and protocols for data protection and security in the cloud, still there are lacking in these models considering their effectiveness [26]. According to Botta, Donato, Persico and Pescap [27], security and privacy preservation issues are significant research challenges in the area of Cloud IoT (i.e. the technology where the cloud and IoT are used as complementary to each other for knowledge management). According to Zhou, Lin, Dong and Cao [28], in the context of m-healthcare cloud computing system, patient self-controllable multi-level privacy-preserving cooperative authentication scheme (PSMPA) needs to be incorporated which will allow the users for self-control for ensuring confidentiality and multi-level privacy. [29] also developed a privacy-preserving auditing protocol called as SecCloud protocol to address the security and privacy challenges. Yet, in another study, [20] have proposed H2Hadoop design for preserving privacy for Big Data in Mobile Cloud. In another study, [14] used a database encryption approach called L-EncDB for protection of data privacy in cloud. Liu, Ning, Xiong and Yang (2015) used shared authority based privacy-preserving authentication (SAPA) protocol for ensuring privacy for the users of collaborative cloud applications. For image retrieval scheme in cloud computing, Xia, Xion, Vasilakos, and Sun [30] used the content-based image retrieval (CBIR) technique to do not reveal sensitive information of the users using cloud. For this purpose, Zhang, Yang, Chen, and Li [31] used the possibilistic c-means algorithm (PCM) to solve the privacy issues in image analysis and knowledge discovery in cloud-computing environment. Likewise, in the e-healthcare cloud computing systems, Cao, Dong and Lin [19] used traceable and revocable multi-authority attribute-based encryption (TR-MABE) approach for addressing the security and privacy issues. Similarly, Yuan and Yu[32] used multi-party network learning scheme for the prevention of revealing of cloud user private information.

4 Conclusion

This review consolidates the related data security and privacy challenges under two techniques used for knowledge discover in cloud environment. Also, the review consolidates the proposals proposed by different experts to address the data security and privacy concerned. The review shows, as the knowledge discovery in the cloud-computing environments entails newer techniques, newer challenges of data security

and privacy are also emerging accordingly and continuously. Both the data security and privacy issues are becoming determinant factors in adopting and applying new techniques in cloud for knowledge discover. It is recommended in the literature that the cloud stakeholders need to work together to minimise the data security issues. In addition, both the hardware and software level measures in the cloud are recommend to ensure the data security, and the subsequent data privacy. On the other hand, various multi-level privacy-preserving protocols are proposed in various studies to address the privacy concerns. It is important to note that the challenges of maintaining data security and privacy exit yet mostly because of the presence of the intrinsic features of Big Data and machine learning linked to security and privacy. Hence, the experts are devising various models and algorithms to balance between the security and privacy aspects and the wellbeing aspects of the techniques used for knowledge discovery. In such situation, cloud computing providers need to continue research and develop of new privacy-focused models for providing knowledge discovery facilities in cloud-computing environments. However, these facilities also need to be convenient beside privacy focus. While data is traveled over the internet, stored and mined in the cloud-computing environments for knowledge discovery using various techniques, the users need to be aware of the concern and challenges regarding privacy and identity protection from the cloud service providers. Otherwise, the practicability of cloud computing for knowledge discovery and popularity of cloud service providers would dramatically decrease. Furthermore, firm-level data collection policy is essential to assess the level of sensitivity of information used in Big data and machine learning. Because of the inadequately addressed security and privacy issues in cloud-computing environments, still researchers are working. However, most of the studies are based on the technical aspects, where the management aspects for privacy-preserving are not focused enough (2015). Finally, the findings show that in the context of cloud-computing environment, security and privacy related issues still require further research as many such issues are yet to be overcome.

References

[1] Derya Birant and Pelin Yıldırım. A framework for data mining and knowledge discovery in cloud computing. In *Data Science and Big Data Computing*, pages 245–267. Springer, 2016.

[2] Edmon Begoli and James Horey. Design principles for effective knowledge discovery from big data. In *2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture*, pages 215–218. IEEE, 2012.

[3] Kun Gao, Qin Wang, and Lifeng Xi. Reduct algorithm based execution times prediction in knowledge discovery cloud computing environment. *Int. Arab J. Inf. Technol.*, 11(3):268–275, 2014.

[4] Domenico Talia. Making knowledge discovery services scalable on clouds for big data mining. In *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, pages 1–4. IEEE, 2015.

[5] Nabil Sultan. Knowledge management in the age of cloud computing and web 2.0: Experiencing the power of disruptive innovations. *International journal of information management*, 33(1):160–165, 2013.

[6] Nir Kshetri. Big data impact on privacy, security and consumer welfare. *Telecommunications Policy*, 38(11):1134–1145, 2014.

[7] Christos Stergiou, Kostas E Psannis, Byung-Gyu Kim, and Brij Gupta. Secure integration of iot and cloud computing. *Future Generation Computer Systems*, 78:964–975, 2018.

[8] Maricela-Georgiana Avram. Advantages and challenges of adopting cloud computing from an enterprise perspective. *Procedia Technology*, 12: 529–534, 2014.

[9] Charith Perera, Rajiv Ranjan, Lizhe Wang, Samee U Khan, and Albert Y Zomaya. Big data privacy in the internet of things era. *IT Professional*, 17(3):32–39, 2015.

[10] Munsif Yousef Sokiyna, Musbah J Aqel, and Omar A Naqshbandi. Cloud computing technology algorithms capabilities in managing and processing big data in business organizations: Mapreduce, hadoop, parallel programming. *Journal of Information Technology Management*, 12(3): 100–113, 2020.

[11] Mohamed Almorsy, John Grundy, and Ingo Müller. An analysis of the cloud computing security problem. *arXiv preprint arXiv:1609.01107*, 2016.

[12] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414. IEEE, 2018.

[13] Andreas Holzinger, Matthias Dehmer, and Igor Jurisica. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, 15(6):1–9, 2014.

[14] Ping Li, Jin Li, Zhengan Huang, Tong Li, Chong-Zhi Gao, Siu-Ming Yiu, and Kai Chen. Multi-key privacy-preserving deep learning in cloud computing. *Future Generation Computer Systems*, 74:76–85, 2017.

[15] Ibrahim Arpaci. Antecedents and consequences of

cloud computing adoption in education to achieve knowledge management. *Computers in Human Behavior*, 70:382–390, 2017.

[16] Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu, and Fei Hu. Big data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, 10(1):13–53, 2017.

[17] Jun Zhou, Zhenfu Cao, Xiaolei Dong, and Athanasios V Vasilakos. Security and privacy for cloud-based iot: Challenges. *IEEE Communications Magazine*, 55(1):26–33, 2017.

[18] M Reza Rahimi, Jian Ren, Chi Harold Liu, Athanasios V Vasilakos, and Nalini Venkatasubramanian. Mobile cloud computing: A survey, state of art and future directions. *Mobile Networks and Applications*, 19(2):133–143, 2014.

[19] Ning Cao, Cong Wang, Ming Li, Kui Ren, and Wenjing Lou. Privacy-preserving multi-keyword ranked search over encrypted cloud data. *IEEE Transactions on parallel and distributed systems*, 25(1):222–233, 2013.

[20] T Ramathulasi, C Samba Shiva Reddy, and K Ashok Kumar. Data encryption strategy with privacy-preserving for big data in mobile cloud using h2hadoop. 2018.

[21] Zhangjie Fu, Jiangang Shu, Jin Wang, Yuling Liu, and Sungyoung Lee. Privacy-preserving smart similarity search based on simhash over encrypted data in cloud computing. *Journal of Internet Technology*, 16(3):453–460, 2015.

[22] Munsif Sokiyna and Musbah Aqel. The role of e-business applications software in driving operational excellence: Impact of departments collaboration using sustainable software. *Sustainable Computing: Informatics and Systems*, 28:100445, 2020.

[23] Joseph K Liu, Man Ho Au, Xinyi Huang, Rongxing Lu, and Jin Li. Fine-grained two-factor access control for web-based cloud computing services. *IEEE Transactions on Information Forensics and Security*, 11(3):484–497, 2015.

[24] Xun Yi, Fang-Yu Rao, Elisa Bertino, and Athman Bouguettaya. Privacy-preserving association rule mining in cloud computing. In *Proceedings of the 10th ACM symposium on information, computer and communications security*, pages 439–450, 2015.

[25] Yong Yu, Man Ho Au, Giuseppe Ateniese, Xinyi Huang, Willy Susilo, Yuanshun Dai, and Geyong Min. Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. *IEEE Transactions on Information Forensics and Security*, 12(4):767–778, 2016.

[26] Yunchuan Sun, Junsheng Zhang, Yongping Xiong, and Guangyu Zhu. Data security and privacy in cloud computing. *International Journal of Distributed Sensor Networks*, 10(7):190903, 2014.

[27] Alessio Botta, Walter De Donato, Valerio Persico, and Antonio Pescapé. Integration of cloud computing and internet of things: a survey. *Future generation computer systems*, 56:684–700, 2016.

[28] Jun Zhou, Zhenfu Cao, Xiaolei Dong, and Xiaodong Lin. Tr-mabe: White-box traceable and revocable multi-authority attribute-based encryption and its applications to multi-level privacy-preserving e-healthcare cloud computing systems. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2398–2406. IEEE, 2015.

[29] Lifei Wei, Haojin Zhu, Zhenfu Cao, Xiaolei Dong, Weiwei Jia, Yunlu Chen, and Athanasios V Vasilakos. Security and privacy for storage and computation in cloud computing. *Information sciences*, 258:371–386, 2014.

[30] Zhihua Xia, Neal N Xiong, Athanasios V Vasilakos, and Xingming Sun. Epcbir: An efficient and privacy-preserving content-based image retrieval scheme in cloud computing. *Information Sciences*, 387:195–204, 2017.

[31] Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li. Pphopcm: Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing. *IEEE Transactions on Big Data*, 2017.

[32] Jiawei Yuan and Shucheng Yu. Privacy preserving back-propagation neural network learning made practical with cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, 25(1): 212–221, 2013.

Najah Kalifah Almazmomi received her Ph.D. degree in management from School of Management (MIS), La Trobe University, 2016, Melbourne VIC, Australia. Currently, she is an assistant professor in the department of Management of Information System, College of Business, University of Jeddah, Saudi Arabia.