IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SURVEY: PRIVACY-PRESERVING VERIFIABLE SET OPERATION IN BIG DATA FOR CLOUD-ASSISTED MOBILE CROWDSOURCING

[1]T Asha Latha, [2]Niteesha Sharma, [3] N Naga Lakshmi

[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor

[1]Information Technology,

[1]Anurag University, India

Abstract: The usage of smartphone made possible for many real-time applications such as mobile crowdsourcing, by which the task owner can crowdsource the data from the smartphone users. There are many challenging problems such as data aggregation, data analysis and data collection faced by the resource constrained requester, when the data volume is extremely large i.e., Big Data (BD). The process of filtering redundant data and preprocessing raw data from the Big Data Analysis (BDA) by using set operations such as union, complementation and intersection. The most promising way for solving the BDA issues is the cloud-assisted approaches in terms of limited number of resources like computation and storage resources. If the privacy of the smartphone user's sensing data and identity are preserved well, the users will be ready to participate in the untrusted cloud. But due to the security issues faced by the cloud, the smartphone users are not willing to share their private data with the task owner. This review paper evaluates the researches done on the BDA and also auditing the major issues face by several techniques. The researchers can give the better solution for the current problems faced by using this procedure in the BDA

Index Terms - Mobile Crowdsourcing, Smartphone users, Task owner, Big data Analysis, Sensing Data

I. INTRODUCTION

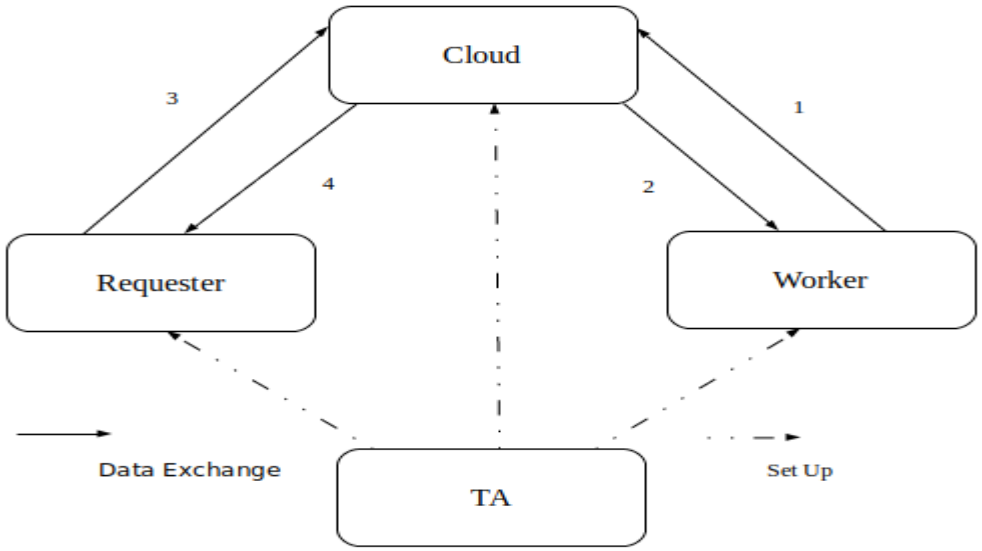
II. In the recent years, Internet has become a fundamental component of everyday social and business life for billions of people around the world. Internet users exploit on a daily-basis a vast range of web-based applications, ranging from on-line shopping and banking to social networks [1]. The task is difficult to store and manage these high structural data sets while considering security and privacy issues. Organizations, like hospitals, generate a large volume of input where maximum of the generated information is helpful only when they are share [2]. The implementation of strong security for BD size which is greater than 1 Terra Byte to the data center being approached is one of most interesting topics. Hence, there are lot of solutions are available in data centers for resolving the data security. In common, the collection of large amounts of data sets with different types is defined as BD andso the process becomes difficult to proceed by using popular data processing algorithms and platforms [3]. The environment of data provisions such as sensor networks, satellite and streaming machines, social networks and high throughput instruments has increased nowdays which produces huge size of data. In many applications such as natural resources, social networking, health care, education, etc., BD is used for providing the data to the above applications. The BD information should be imported in high secure manner due to the process of hacking the information. The techniques such as encryption, honeypot detection and logging must be necessary for securing the BD [4]. Considering the complexity and heterogeneity of data streams and sources, performing BDA is a challenging task [5].

III. The process of detecting the hackers and preventing the information from the malicious intruders and advanced threats, BDA is used [6]. The Data pre-processing, data transformation, data mining and pattern evaluation and presentation are the basic steps to achieve the useful information in BD [7]. The organization for exploring, developing and realizing real-time commercial insights, BD excitements are used in several business touch points like consumers, productivity and shareholder wealth maximization. BD research is multi-dimensional and facilitates the growth of scientific discovery and innovation [8]. Applying BD offers significant benefits for individuals and society, but also raises serious concerns about several information security risks like data security, governance, and privacy. Failure to understand the BD risks will defeat the purpose of the value generation of BD and also will cause the organization vulnerable to critical and irrevocable data losses [9]. Thus, we can control the security and the privacy of BD which may be either sensitive or secret information [10]. In this paper, review on the BD has been done in order to analysis the performance and concerns of developed methodologies. This process motivates the researcher's for further research work in BDA.

2. System model:

The system model mainly consists of four objects such as the mobile workers (W), the requester (R), the cloud (C) and the trusted authority (TA). The basic structure of the cloud architecture is described as in the figure 1.,

Figure 1. Structure of the cloud architecture



where,

1. Datasets,
2. Broadcast,
3. Task Delegation,
4. Results Retrieval

- **Trust Authority:**
  TA is responsible for generating the system parameters which includes registering workers, requesters and the cloud, generating public parameters, and distributing keys, and maintaining the system [11]. TA is also tasked to create the public/secret key pair for the data cloud and data users. Both the cloud provider and the user trusted the trust auditor. While auditing the data on user, the TA minimizes the computational data pressure and TA may be offline unless a dispute arises [12].
- **Requester:**
  The requester needs to send a request for the worker to obtain the intersection set of the data of workers. The requester control message is sent by the initial requester node first and it comprises the following two main components such as the requester ID, the requester unique identifier and the service request, that is, what the node is specifically requesting [13]. The requester will delegate storage and the computation task due to the shortage on the storage and capability of computation of user to the cloud.
- **Cloud:**
  The public cloud has almost unlimited storage and computing power to undertake the storage task and respond on data retrieval requests [14]. The delegation requests are send by the requesters whereas, the mobile workers send the encrypted data to the cloud. The intersection set for the requester is computed by the cloud [15]. The cloud also needs to provide some proof information to prove the correctness of the result.
- **Mobile Workers:**
  The workers are willing to contribute data from the smartphones to the requester's tasks is referred to as Mobile Workers. The generation of data set by themselves and the workers can encrypt the data before sending to the cloud. When a certain mobile workers wants to issue search queries over encrypted cloud data, he needs to submit a search token generated by his specified keywords to cloud service provider [16]. Then the cloud returns relevant search results if and only if the attributes satisfy the specified access policy.

2.1. Security Model:

In our security model, the TA is fully trusted and will not be breached by any adversary. In analysing the privacy of data, a worker's data set should be kept confidential from other workers and cloud. The requester can able to check the precision of the computation result received from the cloud [17]. The Cloud are honest-but-curious, which describes that they will strictly follow the predefined protocol, but they may try to know other's sensitive data based on available information [18]. In [32], scheme may fail to fight against the denial-of-service attacks when numbers of malicious workers send requests to the cloud with dummy crowdsourcing data.

2.2. Design Objective:

The ultimate goals of research in [19] on engineering design is to improve communication in the architecture process. In traditional collaborative design settings, communication can be seen as a straightforward process with a linear sequence of design phases. Because of the social media usage in Cloud-based design (CBD) settings, design communication can be improved through multiple information channels (e.g., social network sites and product review sites) in which information flow can take place in multiple directions. The cloud storage usually has a storage hierarchy such as the memory (primary storage) and disk (secondary storage). The execution of a Proof of Ownership (POW) scheme might require the user and cloud to access the file stored in the disk multiple times. The server might also lead to keep the verification object in either the memory or the disk to verify the user's claim. The above all might result in a large volume of I/O delay because of the access time gap between the memory and disk. In [20], the corruption of a file hash is focused by the method for gaining the ownership of the file and the main aim of the model is to design a scheme of POW with minimum performance overhead.

3. Literature Review:

Several techniques are suggested by researchers in the BDA. In this scenario, brief evaluations of some important contributions to the existing techniques are presented.

Author	Methodology Employed	Dataset	Advantage	Limitation	Performance measure
V. S. Thiyagarajan, and A. Ayyasamy, [21]	Variation step size firefly algorithm and MapReduce framework	Census-Income (KDD) dataset	The proposed algorithm improves the maximum accuracy, privacy preserving and scalability when compared to the existing Neural Network	The method need more information for testing the effectiveness of the method. Suppose, if the fireflies are holding the worst fitness value, the threshold value will affect the performance.	Mean Magnitude of Relative Error (MMRE) and Accuracy
H. Rong, *et al.*, [22]	Privacy-preserving K-Nearest Neighbor (KNN)	Wine Quality dataset from UCI Machine Learning Repository	The paper ensures the confidentiality of data, KNN query, results and access pattern with small computational and communication costs.	The proposed method was not efficient in large scale database	Cloud Computation Time, communication cost, and Data owner computation
P. Hu, *et al.*, [23]	Fog computing based face identification and resolution framework	Georgia Tech (GT) face, Caltech face and BioID face databases	The framework can solve the issues of confidentiality, integrity and availability	While ensuring the privacy of the system, the paper increases only a little computation and communication overhead	Response time and Amount of network transmission
Z. Wang, *et al.*, [24]	Ciphertext-Policy Attribute-Based Encryption (CP-ABE) and Key-Policy Attribute-Based Encryption (KP-ABE)	Unknown	The theoretical and experimental results showed that the model is a good secure solutions for the volume and velocity of BD.	The encryption process is executed with constrained resources is not an automated process, hence failure may occur in the method causes poor efficiency	Encrypt and decrypt time, Leakage bound and Leakage model
D. He, *et al.*, [25]	Privacy-preserving CertificateLess Provable Data Possession (PP-CLPDP)	Unknown	The set of experimental results shows that the PP-CLPDP scheme providing better data integrity for BD	The proposed approach is not applicable for real-world cloud datasets	Computation cost and communication cost
			The proposed scheme guarantees user's anonymity, so the security awareness information is generated by the bit	The information sent and received between users and servers are not integrate and unable to manage	Server computation, user computation and communicationcomplexity

Jeong Yoon-Su, and Seung-Soo Shin, [26]	Service Management Scheme Protocol	Unknown	sequence which is not easily exposed to third party	the stratified properties by the proposed method	
P. Li, *et al.*, [27]	Oblivious RAM (ORAM)	Unknown	The method achieved a load-balanced storage system by studying data-placement problem with increased responsiveness and availability	The complexity of the ORAM model leads to the poor performance	Maximum Access Rate and Execution time
H. A. Al Hamid, *et al.*, [28]	Tri-party one-round authenticated key agreement protocol	Unknown	The original multimedia data is more secure by setting the default value of the decoy data.	The method uses fog computing technique to create decoys files but these files can be established for only minimal user intervention	Computational cost and communication overhead
Y. Yang, *et al.*, [29]	Privacy-preserving healthcare BD storage and self-adaptive access control system with smart deduplication	Unknown	The proposed model achieves versatile useful functions and efficient by means of the storage and computation costs	The method proposed a password based break-glass key algorithm for encounter some emergency situation of patient. But the process takes more time to encrypt the secured data	The performance measures includes functionality and costs of computation.
Y. Yang, *et al.*, [30]	A keyword match based policy update mechanism	Unknown	The group of keys are distributed to the medical nodes by the users without any interaction in an authenticated way.	The encrypted data can be updated by the user with the help of keyword match based update process. Suppose, the user forget the keyword, they are unable to update in the medical BD system	Computation overhead, communication overhead, transmission efficiency and computation efficiency,
Y. Yan, *et al.*, [31]	A hierarchical differential privacy hybrid decomposition algorithm	Unknown	The algorithm has good effect in improving the accuracy of regional counting queries and also have less complexity in computation	The adaptive grids decomposition method takes the location BD publishing process into only one snapshot of data, which cannot be applied to the publishing of stream data of location with high speed of arrivals and fast dynamic changing	Relative Error and accuracy of range queries
G. Zhuo, *et al.*, [32]	Privacy-preserving Verifiable Set Operation	Unknown	The paper preserved the data of workers and identity privacy and the set operation results are verified by the requester	If the workers increases, the cost of correctness verification also increases which leads to the damage for battery- limited devices.	Computational cost and cost reduction

K. Fan, *et al.*, [33]	Secure Key management scheme	Unknown	The method can ensure the secure and privacy of the data by which the unauthorized users cannot decrypt the user's data.	The model increases the time for encryption, the key distribution for security of keys. The process of encrypting and decrypting the data will occur in the server-side only.	The time for key generation and the time for encryption and decryption performance
H. He, *et al.*, [34]	Big-data aided hybrid relay selection scheme	Unknown	The extra overhead caused by the increased number of relays is resolvable by using this model	The hybrid technique yields poor performance when the amount of relays becomes large under the scenario that the Eavesdropper's instantaneous Channel state information	Security outage probability
C. Lin, *et al.*, [35]	Privacy protection scheme for sensitive BD by using Haar Wavelet Transformation method	Unknown	The tree structure is greatly reduced the calculation overhead which is demonstrated by the paper and also preserving the privacy for users.	When the value of noise is increased, the availability of data will be improved, but the security and privacy protection will be affected.	Frequency of original data, added noise data and deleted noise data.

4. Conclusion

Data over the internet has been rapidly increasing day by day. Nowdays, the organizations who are all having the large datasets, they will automatically mine the useful information from the massive data moreover this is considered as a common concern. Here, the privacy preserving is has showed as a highly significant challenge encountered by the data mining domain, as it is extremely hard to conserve the confidential documents of the clients. While ensuring in the meantime delicate data, the process of segregating the significant learning from a more information is the main aim of the privacy preserving calculations. The IT advancements and user-friendly global services can be globalized and offered by a single click by means of cloud applications like the BD. The innovative feature of the BD makes it endearing in view of the cost-conscious nature of storage and processing of the relative datasets. This review paper gives an overview of BD in privacy preserving and also evaluates the developed methodologies by means of advantage, limitation and performance measure. Additionally, reviews several commonly utilized multimodal datasets, and empirically evaluates the concerns faced by the methodologies. Still, there is much work to be done on privacy preserving in BD for delivering better privacy. This review paper will help readers to understand the state-of-the-art in privacy preserving and motivate more meaningful works.

Reference

[1] Kobusińska, Anna, Kamil Pawluczuk, and Jerzy Brzeziński. "Big Data fingerprinting information analytics for sustainability," *Future Generation Computer Systems* , 2018.

[2] El Ouazzani, Zakariae, and Hanan El Bakkali. "A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k." *Procedia Computer Science*, vol. 127, pp. 52-59, 2018.

[3] Manogaran, Gunasekaran, and Daphne Lopez. "A Gaussian process based big data processing framework in cluster computing environment." *Cluster Computing* , pp. 116, 2017.

[4] Manogaran, Gunasekaran, Chandu Thota, and M. Vijay Kumar. "MetaCloudDataStorage architecture for big data security in cloud computing." *Procedia Computer Science* , vol. 87, pp. 128-133, 2016.

[5] M. H. ur Rehman, E. Ahmed, I. Yaqoob, I. A. T. Hashem, M. Imran, and S. Ahmad, "Big Data Analytics in Industrial IoT Using a Concentric Computing Model", *IEEE Communications Magazine*, vol. 56, no. 2, pp. 37-43, 2018.

[6] Pena Pedro A., Dilip Sarkar, and Parul Maheshwari. "A big-data centric framework for smart systems in the world of internet of everything." *Computational Science and Computational Intelligence (CSCI), 2015 International Conference on*. IEEE, 2015.

[7] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: privacy and data mining," *IEEE Access*, vol. 2, pp. 1149-1176, 2014.

[8] Bharathi, S. Vijayakumar. "Prioritizing and ranking the big data information security risk spectrum," *Global Journal of Flexible Systems Management* vol. 18, no. 3, pp. 183-201, 2017.

[9] D. K. Venkatachalapathy, V. S. Thiyagarajan, A. Ayyasamy, and K. Ranjani, "Big data with cloud virtualization for effective resource handling," *International Journal of control Theory and Applications*, vol. 9, no. 1, 2016.

[10] Thayananthan, and Aiiad Albeshri. "Big data security issues based on quantum cryptography and privacy with authentication for mobile data center." *Procedia Computer Science* vol. 50, pp. 149-156, 2015.

[11] Y. Miao, J. Ma, X. Liu, X. Li, Q. Jiang, and J. Zhang, "Attribute-based keyword search over hierarchical data in cloud computing," *IEEE Transactions on Services Computing*, 2017.
[12] Wei, Jinqiao, Ying Wang, and Xiaoxue Ma. "Text image authenticating algorithm based on MD5-hash function and Henon map." Ninth International Conference on Digital Image Processing (ICDIP 2017). Vol. 10420. International Society for Optics and Photonics, 2017.
[13] B. M. Silva, J. J. Rodrigues, F. Canelo, I. C. Lopes, and L. Zhou, "A data encryption solution for mobile health apps in cooperation environments", *Journal of medical Internet research*, vol. 15, no. 4, 2013.
[14] Y. Yang, X. Liu, R. H. Deng, and Y. Li, "Lightweight sharable and traceable secure mobile health system," *IEEE Transactions on Dependable and Secure Computing*, 2017.
[15] M. Li, J. Weng, A. Yang, W. Lu, Y. Zhang, L. Hou, and J. Liu, "*CrowdBC: A Blockchain-based Decentralized Framework for Crowdsourcing",* IACR Cryptology ePrint Archive, pp. 444, 2017.
[16] Y. Miao, J. Ma, X. Liu, X. Li, Z. Liu, and H. Li, "Practical Attribute-Based Multi-keyword Search Scheme in Mobile Crowdsourcing", *IEEE Internet of Things Journal*, 2017.
[17] Q. Meng, J. Ma, K. Chen, Y. Miao, and T. Yang, "Comparable Encryption Scheme over Encrypted Cloud Data in Internet of Everything", *Security and Communication Networks*, 2017.
[18] Zhuo Gaoqiang, and Huihui Yang. "Privacy-preserving context-aware friend discovery based on mobile sensing." *Consumer Electronics (ICCE), 2018 IEEE International Conference on*. IEEE, 2018.
[19] D. Wu, D. W. Rosen, L. Wang, and D. Schaefer, "Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation", *Computer-Aided Design*, vol. 59, pp. 1-14, 2015.
[20] Yu, Chia-Mu, Chi-Yuan Chen, and Han-Chieh Chao, "Proof of ownership in deduplicated cloud storage with mobile device efficiency," *IEEE Network*, vol. 29, no. 2, pp. 51-55, 2015.
[21] V. S. Thiyagarajan, and A. Ayyasamy. "Privacy Preserving Over Big Data Through VSSFA and MapReduce Framework in Cloud Environment." *Wireless Personal Communications*, vol. 97, no. 4. pp. 6239-6263, 2017.
[22] H. Rong, H. M. Wang, J. Liu, and M. Xian, "Privacy-preserving k-nearest neighbor computation in multiple cloud environments," *IEEE Access*, , vol. 4, pp. 9589-9603, 2016.
[23] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, "Security and privacy preservation scheme of face identification and resolution framework using fog computing in internet of things," *IEEE Internet of Things Journal*, vol, 4, no. 5, pp. 1143-1155, 2017.
[24] Z. Wang, C. Cao, N. Yang, and V. Chang, "ABE with improved auxiliary input for big data security," *Journal of Computer and System Sciences*, , vol. 89, pp. 41-50, 2017.
[25] D. He, N. Kumar, H. Wang, L. Wang, and K. K. R. Choo, "Privacy-preserving certificateless provable data possession scheme for big data storage on cloud," *Applied Mathematics and Computation*, vol. 314, pp. 31-43, 2017.
[26] Jeong Yoon-Su, and Seung-Soo Shin, "An efficient authentication scheme to protect user privacy in seamless big data services," *Wireless Personal Communications* , vol. 86, no. 1, pp. 7-19, 2016.
[27] P. Li, S. Guo, T. Miyazaki, M. Xie, J. Hu, and W. Zhuang, "Privacy-preserving access to big data in the cloud," *IEEE Cloud Computing*, , vol. 3, no. 5, pp. 34-42, 2016.
[28] H. A. Al Hamid, S. M. M. Rahman, M. S. Hossain, A. Almogren, and A. Alamri, "A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With Pairing-Based Cryptography," *IEEE Access*, , vol. 5, pp. 22313-22328, 2017.
[29] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, "Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system," *Information Sciences*, 2018.
[30] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, "Privacy-preserving fusion of IoT and big data for e-health," *Future Generation Computer Systems*, 2018.
[31] Y. Yan, Xiaohong Hao, and Lianxiu Zhang, "Hierarchical differential privacy hybrid decomposition algorithm for location big data," *Cluster Computing*, pp.1-12, 2018.
[32] G. Zhuo, Q. Jia, L. Guo, M. Li, and P. Li, "Privacy-preserving verifiable set operation in big data for cloud-assisted mobile crowdsourcing," *IEEE Internet of Things Journal*, , vol. 4, no. 2, pp. 572-582, 2017.
[33] K. Fan, S. Lou, R. Su, H. Li, and Y. Yang, "Secure and private key management scheme in big data networking," *Peer-to-Peer Networking and Applications*, pp. 1-8, 2017.
[34] H. He, P. Ren, Q. Du, L. Sun, and Y. Wang, "Enhancing physical-layer security via big-data-aided hybrid relay selection," *Journal of Communications and Information Networks*, vol. 2, no. 1, pp. 97-110, 2017.
[35] C. Lin, P. Wang, H. Song, Y. Zhou, Q. Liu, and G. Wu, "A differential privacy protection scheme for sensitive big data in body sensor networks," *Annals of Telecommunications*, vol. 71, no. 9-10, pp. 465-475, 2016.