ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Privacy Preservation and Copyright Protection for Big Data in Hadoop Environment: A Survey

Dhanashri Varute[1], Arti Mohanpurkar[2]

Department of Computer Engineering, D.Y.Patil's SOET, Lohegaon, Pune, India

ABSTRACT: Nowadays, cloud computing and big data are two intrusive trends, providing multiple scope to the IT industry and research communities while pretense obvious challenges on them as well. As big data consist of tremendous amount of personally identifiable information, user privacy and data security are the major concerns and massive challenges in the big data. Concept of a big data may not be widely accepted if privacy and security are not well addressed. Data anonymization and encryption are two widely adopted ways to prevent the information from privacy breach. But for the data which is processed and shared frequently, encryption is not suitable. Anonymizing big data and managing number of anonymized datasets is still challenging for traditional anonymization approaches. Big data could be structured, semi-structured or unstructured, which adds more challenges. This paper summarizes all techniques previously used to provide privacy as well as security to information of big data.

KEYWORDS: Privacy Preservation, Map Reduce, Anonymization, data management, Cloud Computing, Big Data, Fingerprint technique.

I. INTRODUCTION

In present day scenario, cloud has become an inevitable need for most of Information Technology (IT) organizations. The word cloud meant for particular divergent IT infrastructure and environment which is specifically developed for fulfilling the outline of remote access to the scalable and managed resources. In cloud computing the data resources are shared instead of obligating the personal server applications. The infrastructure of cloud facilitates the allied users to access and utilize the resources according to their requirement in the real time applications. For IT organizations dealing with cloud computing, applications such as data retrieval, data portability and data storage have become important need. Considering the requirement, the IT development and user oriented global services are globalized and delivered to single click by cloud application like Big Data. We are living in era where we experience a true big data explosion, a bang in databases and database technologies. The applications such as World Wide Web, mobile computing, and wireless technologies generate a tremendous amount of data called big data. To preserve privacy of data different mechanisms have been proposed and developed in the past years. The great challenge is to keep balance between data utility and data privacy. Nowadays there is a great demand for optimum data access and resource utilization system with minimum latency and complexity. For this there must be a fare management optimization in processing units of Hadoop. Big data raise the concerns about tracking and profiling of consumers and people as it expands to all domains. The nature of big data introduces an immense challenge to privacy preservation. In the past days, big data access was limited only to the governments and large enterprises, but nowadays, big data is accessible to everyone. If this data is accessed and analyzed properly, then it enable to offer a better understanding of human behaviors which helps the organizations to improve their business and contributes to the field of global development too. An unauthorized user access and misuse of information collected from the users violates the protection of individual privacy. Since multiple sources are involved, the risk of privacy gets doubled, so protection of privacy in big data is a fast growing field of research. So these violations lead to unconvincing knowledge mining results and user reluctance in data sharing. Such issues somewhat contributes in the Privacy-Preserving Data Publishing as well as security problems. Privacy protection concept states that accessing of published data must not allow the unwanted users to identify anything about the targeted individuals and security states that the published data must be protected

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

from copyright. To accomplish the optimum privacy preservation, approach of data anonymization plays an important role. In case of large amount of information in Big Data applications, to accomplish optimum function, scalability and efficiency, the system paradigms move towards Map Reduce framework of Hadoop model. To accomplish security of data fingerprint techniques plays vital role.

III. LITERATURE SURVEY

Authors [1] presented an existing privacy preserving mechanisms in the various life cycles of big data like data generation, data storage and data processing (anonymization techniques such as generalization, suppression, anatomization, permutation and perturbation) and various challenges of preserving privacy in big data. These methods are described with respect to the factors of privacy, scalability, time, utility and efficiency. Various risks involved in the anonymization, encryption and storage of data in the cloud are also investigated. When applying these techniques, privacy is protected but the data may lose the meaning in the real world and also the utility and significance. Therefore the techniques need to be modified or extended to handle the privacy and security of big data in an efficient manner.

Rampart framework for privacy preservation created in [2] consists of seven procedures as anonymization, reconstruction, modification, provenance, agreement, trade and restriction to prevent outside intrusion. This framework tries to give high priority to maintain the balance between data utility and privacy. But more ways are to be explored to protect privacy against various threats.

Further two solutions – SRA and HPA - for privacy preservation based on differential privacy are presented [3]. The study with real world data shows that the solutions enable accurate data analysis reducing the user privacy risk and data storage. They have considered the fact that a few users may add huge amount of records against the assumption of other approaches that each user contribute only one record. These methods take care to achieve a balance between data loss and privacy.

Privacy preserving problem of big data in the context of hybrid cloud computing is investigated in [4] and presented frameworks such as Airavat, Sedic, Sac-FRAPP and Hyper-1 based on MapReduce from the perspective of scalability, cost and compatibility. Anonymization, encryption and differential privacy are the efficient methods for protecting privacy of data is recorded here. In the final analysis it is shown that the above said frameworks suffers from limitations such as data distortion and no one of them is fully fit for privacy preservation.

A new privacy model with overlapping slicing which duplicates attribute in more than one column is presented in [5]. It is claimed that this model increases privacy and utility of data by achieving correlation among attributes. It can also handle high dimensional data. This solution overcame with the limitations of slicing and anonymization but while reducing attribute disclosure risk increase the chance for identity disclosure risk.

In [6] an approach towards privacy-preserving computing in the big data world is presented which exploits the new challenges of big data in privacy preservation. It defines the general architecture of big data analytics and discovers the privacy requirements in big data. After that, it finds out an efficient and privacy-preserving cosine similarity computing protocol.

MACA [7], first privacy preserving multi-factor authentication system utilizing the features of big data considers both users privacy and usability. In order to protect the privacy, fuzzy hashing and fully homomorphic encryption are used. Facility is provided for modifying the features according to the context. The system must provide options for adding more features which can be configured by the user.

A case study of anonymization in enterprise, enumerating the requirements and implementation details for preserving privacy of big data is presented in [8]. Anonymized data sets must be carefully analyzed, measured and tested whether they are prone to any attacks. Use of Hadoop is suggested to analyze and obtain useful results from the big data. The experiments are conducted on static data set, but it should be extended for real time data sets. This work couldn't definitely conclude that the anonymized data is fully free from any kind of attacks.

Authors [9] analyzed how the differential privacy approach is suitable for big data privacy preservation and presented the different factors that play key role in it. From the various approaches, differential privacy is the best suitable for big data because it is free from the flaws of other approaches and seeks equilibrium between utility and privacy. A framework of perturbation is introduced to achieve the differential privacy.

Authors [10] experimented with real world big data set in cloud from the perspective of defending privacy breaches and to attain high degree of scalability and efficiency. They proposed a proximity privacy model and a scalable two phase

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

clustering approach based on MapReduce performing data parallel computation in cloud to address the issue of privacy. Experiments conducted in a cloud environment named U-cloud shows that the approach improves the capability of defending the privacy attacks. The model partitions the data into clusters where top down anonymization is not applicable and may fall short when encountering big data.

A technique for fingerprinting considering knowledge preservation [11] [12] on numeric relational data ensures that the usability constraints doesn't violates. By optimizing the error which is inserted with use of Particle Swarm Optimization, knowledge preservation is achieved. System is developed by avoiding collusion and it is primary key independent. Fingerprinting technique provides security against ownership theft and helps in traitor tracing.

Traitor identification system [13] embeds fingerprint securely for providing protection to numeric relational databases. Insertion technique proposed [13] reduces time complexity as well as ensures that inserted fingerprint in the form of an error bits leads to minimum distortion. The system performs blind decoding and it is considered as robust against attacks such as tuple insertion and deletion, attribute deletion etc.

III. ASSESSMENT

Various approaches are made available for privacy protection and security for big data in above research works. Though in each approach privacy protection is the main consideration, still privacy of big data needs to be investigated in detail and present approaches should be extended. Misuse of data make privacy breaches and can harm to data subjects.

The approach given in [1] addresses only three characteristics (volume, variety and velocity) of big data but the other characteristics (value, viscosity, virality and variety) should also be considered. Also the generalized data is vulnerable to any attack or not, is not verified using particular tools. Encryption method used has limitations like long execution time, computation overhead, and decryption of data each time for processing. Thus we need an encryption method which can share data without decryption and re-encryption between various parties.

The Rampart framework [2] as a whole appears to be complex and unrealistic. The k-anonymity approach fails to prevent background knowledge. It has limitations such as attribute and record linkage, high execution time and the framework is not implemented with big data.

The framework [3] makes assumption that only few users generate a large amount of data. It does not guarantee on bounding individual data in sample database because system samples the fixed number of records from every user. The system fails in preserving privacy of complex data.

MapReduce along with cloud computing is used for storage of data and privacy preservation in [4]. Already existing mechanisms are mixed here for privacy preservation of data. Disadvantage of this combination is data distortion.

Overlapping slicing [5] has data leakage, risk of identity disclosure as well as redundancy issues because it duplicates the attribute values for protection from the attribute attacks.

In Cosine similarity computing protocol [6] data leakage during the storage, collection and processing is not removed completely also analysis of complex data is not considered.

MACA [7] presents a multi-factor authentication system; which still needs to be extended to ensure privacy preservation of large datasets.

In [8] anonymization and privacy protection techniques are combined together which proves anonymized big data gives benefits in the enterprise environment but it doesn't state clearly that whether the dataset is vulnerable to correlation attack or not.

How much Differential Privacy concept is good for preserving privacy of big data is experimented in [9]. It is carried out for very small amount of data. There is assumption that the adversary is unable to identify targeted data when two different dataset generates similar output.

Scalable MapReduce multidimensional approach [10] must consider privacy of large datasets and give attention on analysis of scalable privacy preservation.

Fingerprint techniques [11] [12] [13] which provides copyright protection and traitor identification are applied in relational databases only.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

IV. CONCLUSION

Hadoop environment is widely used in industry and for research; therefore security and privacy of data are important issues for organizations running on these environments. The size of the data reverberates in all the fields constantly. Analysis of the big data helps in designing and developing the strategies. But privacy and security of big data is a challenging research issue at present. Big data characteristics show that we need prominent techniques for the processing of big data which takes privacy and security as a prime concern. The existing approaches are not efficient and scalable enough to manage anonymization of the increasing amount of big data and protecting privacy as well as providing security. Thus there is need of system which can provide privacy protection to big data as well as provide copyright protection.

REFERENCES

[1] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). "Protection of Big Data Privacy". IEEE Access, 4, pp.1821-1834. doi:10.1109/access.2016.2558446.
[2] Xu, L., Jiang, C., Chen, Y., Wang, J., & Ren, Y. (2016). "A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining". Computer, 49(2), pp.54-62. doi:10.1109/mc.2016.43
[3] Fan, L., & Jin, H. (2015). "A Practical Framework for Privacy-Preserving Data Analytics". Proceedings of the 24th International Conference on World Wide Web - WWW '15. doi:10.1145/2736277.2741122
[4] Al-Aqeeli, S., & Alnifie, G. (2015). "Preserving Privacy in MapReduce Based Clouds: Insight into Frameworks and Approaches". 2015 International Conference on Cloud Computing (ICCC). doi:10.1109/cloudcomp.2015.
[5] Giri S. Suman., & Mukhopdhay Milav. (2014). "Overlapping Slicing with New Privacy Model". International Journal of Scientific Research Publications, Vol 4, Issue6, June 2014.
[6] Lu, R., Zhu, H., Liu, X., Liu, J., & Shao, J. (2014). "Toward efficient and privacy-preserving computing in big data era". IEEE Network,28(4), 46-50. doi:10.1109/mnet.2014.6863131
[7] Liu, W., Uluagac, A. S., & Beyah, R. (2014). "MACA: A privacy-preserving multi-factor cloud authentication system utilizing big data". 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). doi:10.1109/infcomw.2014.6849285
[8] Sedayao, J., Bhardwaj, R., & Gorade, N. (2014). "Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues". 2014 IEEE International Congress on Big Data. doi:10.1109/bigdata.congress.2014.92
[9] Shrivastva, K. M., Rizvi, M., & Singh, S. (2014). "Big Data Privacy Based on Differential Privacy a Hope for Big Data". 2014 International Conference on Computational Intelligence and Communication Networks. doi:10.1109/cicn.2014.167
[10] Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., & Chen, J. (2013). "Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud". IEEE Transactions on Computers IEEE Trans. Computer., 64(8), 2293-2307.doi:10.1109/tc.2014.2360516
[11] Arti Mohanpurkar, Madhuri Joshi, "Fingerprinting Numeric Databases with Information Preservation and Collusion Avoidance", 2015 International Journal of Computer Applications (0975 – 8887) Volume 130 – No.5, 13-18, November2015.
[12] Arti Mohanpurkar, Madhuri Joshi, "Effect of the Novel Anti-Collusion Fingerprinting Scheme on the Knowledge from Numeric Databases", International Journal of Scientific & Engineering Research, Volume 6, Issue 12, 334-338,December-2015 334 ISSN 2229-5518.
[13] Arti Mohanpurkar, Madhuri Joshi, "A Traitor Identification Technique for Numeric Relational Databases with Distortion Minimization and Collusion Avoidance", 2016 International Journal of Ambient Computing and Intelligence Volume 7, Issue 2, 114-137, July-December 2016.