International Journal of Research in Engineering, Science and Management
Volume-2, Issue-3, March-2019
www.ijresm.com | ISSN (Online): 2581-5792

Safe and Secure Data Markets using Merkle Hash Algorithm

K. Keerthana[1], C. Stefie[2], R. Priyadharshini[3], P. Veeralakshmi[4]
[1,2]Student, Dept. of Information Tech., Prince Shri Venkateshwara Padmavathy Engg. College, Chennai, India
[3]Asst. Prof., Dept. of Information Tech., Prince Shri Venkateshwara Padmavathy Engg. College, Chennai, India
[4]Assoc. Prof., Dept. of Info. Tech., Prince Shri Venkateshwara Padmavathy Engg. College, Chennai, India

***Abstract*: Data mining applications have obtained massive growth in today's internet era. Society has developed an insatiable appetite for sharing personal data. Realizing the potential of personal data's economic value in decision making and user experience enhancement, several open information platforms have emerged to enable person-specific data to be interchanged on the Internet. It eventually leads to a serious threat to the security of an individual's personal and sensitive information. Raw data of millions of users ate collected by the service provider's through data contributors and they a shared with the data consumers. For example social enterprise API platform, collects social media data from users, mines deep insights into customized audiences, and provides data analysis solutions to more than 95% of the Fortune. Existing systems use Privacy Preserving data mining(PPDM) modifies the data without compromising the security of the sensitive information contained in the data. In this paper, we have proposed a new technique called and pseudo identity. The hashing is done using Merkle Hash algorithm.**

***Keywords*: Data markets Privacy, Data truthfulness, Homomorphic algorithm, Merkle Hashing algorithm**

1. Introduction

In the era of big data, society has developed an insatiable appetite for sharing market data. Realizing the potential of personal data's economic value in decision making and user experience enhancement, several open information platforms have emerged to enable marketing specific data to be exchanged on the Internet. To integrate truthfulness and privacy preservation in a practical data market, there are four major challenges. The first and the thorniest design challenge is that verifying the truthfulness of data collection and preserving the privacy seem to be contradictory objectives. Ensuring the truthfulness of data collection allows the data consumers to verify the validities of data contributor's identities and the content of raw data whereas privacy preservation tends to prevent them from learning these confidential contents. Yet, another challenge comes from data processing, which makes verifying the truthfulness of data collection even harder. Nowadays, more and more data markets provide data services rather than directly offering raw data. The third challenge lies in how to guarantee the truthfulness of data processing, under the information asymmetry between the data consumer and the service provider due to data confidentiality. Last but not least, the fourth design challenge is the efficiency requirement of data markets, especially for data acquisition, i.e., the service provider should be able to collect the from a large number of data contributors with low latency. It can used in various real time data markets [1], [2].

2. Related work

The motivation is through various related woks done by various people all over the world. They are as follows: Firstly, Data markets in the cloud, Cloud-computing is transforming many aspects of data management. Most recently, the cloud is seeing the emergence of digital markets for data and associated services. But it could not provide data security and it does not have consistency. Secondly, Anonymous Publication of Sensitive Transactional Data with Low Information Loss, where the objective was to protect the data for information loss and it will provide data security and used greedy cut approach but it cannot handle large datasets. Thirdly, Privacy and Accountability for Location-based Aggregate Statistics, where the objective was to guarantee location privacy and privacy preservation and used strict privacy protocol which gives only guarantees it but does not provide any verification. Fourth, Generating Privacy Recommendation Using Elgamal Homomorphic Encryption, where the objective was to protect the private data from the service provider with functionality of the system is preserved. It used Elgamal Homomorphic Encryption which handles only static inputs and cannot handle dynamic system. Fifth, Privacy Preserving Public Auditing for Secure Cloud Storage, where the objective was to provide a secure cloud storage system supporting privacy-preserving public auditing. The concept use was Third party auditor to monitor the integrity and privacy but the third party auditor cannot process simultaneously but only in batch.

3. System design

In this paper, by jointly considering above four challenges, we propose TPDM, which achieves both Truthfulness and Privacy preservation in Data Markets. TPDM first exploits partially homomorphic encryption to construct a ciphertext space, which enables the service provider to launch data services and the data consumers to verify the correctness and completeness of data processing results, while maintaining data confidentiality. This appealing property can convince data consumers that the service provider has truthfully collected data. The first efficient secure scheme TPDM for data markets, which simultaneously guarantees data truthfulness and privacy preservation. In TPDM, the data contributors have to truthfully submit their own data, but cannot impersonate others. Besides, the service provider is enforced to truthfully collect and process data. Furthermore, both the personally identifiable information and the sensitive raw data of data contributors are well protected using pseudo identity. In addition, we use homomorphic algorithm and Merkle hash algorithm were the data modification can be identified using the root hash generated by the Merkle hash algorithm.

A. System model

There are four actor's data contributors, data collector, dealers, client. The data contributors contribute their data and it is stored in csv format that is, Price of the products

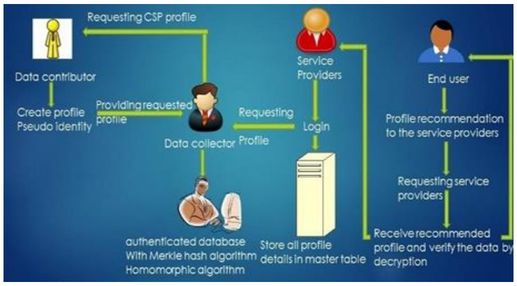
Requesting CSP profile
Data contributor
Create profile
Pseudo identity
Providing requested profile
Data collector
Requesting
Profile
Service Providers
Login
End user
Profile recommendation to the service providers
Requesting service providers
authenticated database
With Merkle hash algorithm
Homomorphic algorithm
Store all profile details in master table
Receive recommended profile and verify the data by decryption

Fig. 1. Block diagram

The data contributors are provided with the pseudo identity to preserve their privacy. The data is then sent to the data collector who dumps the data and manages them and stores the encrypted database. The homomorphic algorithm is used to encrypt the database. The dealers will purchase the database using online payment. When the dealer buy the database simultaneously the hash value is generated in the backend using the Merkle Hash Algorithm. The end user can check the dealer is genuine or not. If the dealer has made any change then the hash value changes and the warning is displayed to the client. So that the client can purchase from the genuine.

As the dealers may change the data for their own personal needs or financial needs or even to attain their targets. Fig. 1, contains the block diagram in which all the operations are clearly explained.

B. Algorithm

1) Merkle hash tree algorithm

Merkle tree is a tree where n hash values results in a single hash, in which every leaf node is labelled with the hash of a data block, and every non-leaf node is labelled with the Cryptographic hash of the labels of its child nodes. Hash trees allow efficient and secure verification of the contents of large data structures. Hash trees are a generalization of hash list and hash chains. The top node is the root hash. The leaf node is converted into hash value using Sha algorithm and each leaf is concatenated to its neighbor node. If there is odd number of leaf it automatically creates a duplicate node. Fig. 2, shows the implementation of the Merkle Hash Tree Algorithm.

Steps:

- The leaf node is hashed using Sha algorithm, which forms the first layer.
- Then the hash values are concatenated with its neighbour.
- This process is done until a single hash is generated.

Pseudocode:

```
1. Set leaf=0
2. Output:
        • Compute and output leaf with LEAFCAL(l).
        • For each h ε [0,H-1] output {$authori_h$}.
3. Refresh authori nodes:
   • For h such that $2^h$ divides l+1.
   • Set $authori_h$ be the sole node value in $stack_h$.
   • Set beginningnode = (l+2h+1) ⊕2h.
   • $stack_h$.initia(firstnode,h).
4. Build stacks:
   • For all h ε [0,H-1].
   • $Stack_h$.update(2).
5. Loop:
   • Set l=l+1.
   • If $l < 2^H$ , go to step 2.
```

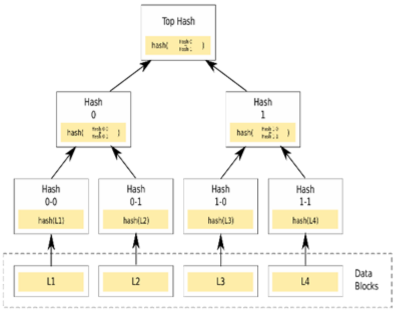

Fig. 2. Merkle hash tree

2) Partially homomorphic algorithm

A Partially Homomorphic encryption is a form of encryption that allows computations on ciphertexts, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the plaintext. The data is encrypted using RSA and it is concatenated to form the encrypted text.

In this system we use this for encrypted the database. Typical database encryption leaves the database encrypted at rest, but when queries are performed the data must be decrypted in order to be parsed. Homomorphic encryption schemes have been devised such that database queries can run against ciphertext data directly.

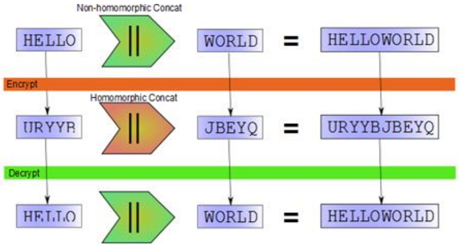

Fig. 3. Partially homomorphic algorithm

The Partially homomorphic algorithm property is given by:

$$\varepsilon(x1).\varepsilon(x2) = x^{e}_{1}\ x^{e}_{2} \bmod m = (x1,x2)^{e} \bmod m = \varepsilon(x1,x2).$$

C. *Data processing using Tpdm*

The analysis of truthfulness of data processing from two aspects, i.e., correctness, completeness. Correctness- TPDM promises the truthfulness of data collection, which is the premise of a correct data service. Then, given a truthfully collected dataset, the data consumer can evaluate over the n contributor’s data sources, which is non-anonymous with the original data processing under the partially homomorphic properties. Completeness- The design provides the property of completeness by guaranteeing the correctness of a, b, and n, which are the numbers of all, correct (authorized), and data contributors, respectively: Initially, the service provider cannot avoid a data contributor’s real data as his wish. The reason is that if the data contributor has submitted his encrypted raw data, without finding his pseudo identity, he would obtain zero benefit for data contribution. Therefore, he has award to report data missing to the registration center, which in turn ensures the correctness of b. Second, it is considered that the service provider compromises the number of valid data contributors a in two ways: one is to put a valid data contributor’s pseudo identity into the warning list; the other is to put an invalid pseudo identity into the good list.

D. *Analysis of Tpdm*

The practical feasibility of TPDM in current data markets. First, to the best of our knowledge, the current applications in real-world data markets, e.g., Microsoft Azure Marketplace have not provided the security guarantees studied in the TPDM framework. The profile matching service, when supporting as many as 1 million data contributors, the computation overhead at the service provider is 0.930s per matching with 10 evaluating attributes in each profile. Besides, for the data distribution service, when supporting 10000 data contributors and 10 random variables, the computation overhead at the service provider is 145s in total. Furthermore, the most time-consuming part of the service provider in TPDM is the computation on encrypted data due to data confidentiality.

4. Result

In this system, after implementing the TPDM approaches, the user or the client can verify the genuine of the dealer and also the privacy can be preserved. Also if any modification is being done then is can be found out before the purchase as we have used the Merkle Hash Algorithm. The privacy of the data contributors is achieved using the pseudo identity given to each of them. The data is been protected using partially homomorphic algorithm. As a result, we have achieved the truthfulness of the data and privacy preservation of the data contributors.

5. Conclusion

The TPDM system provides data truthfulness and privacy preservation in data markets. We use Merkle hash tree algorithm to find whether the data is true or fake. And the partially homomorphic algorithm is used to encrypt the database. The privacy is provided by using pseudo identity. This system overcomes the disadvantages of the existing system by providing security, privacy and truthful data. It can also be used in online shopping websites, also widely in e-commerce industry, stock markets, political surveys, sensitive surveys and so on.

6. Future enhancement

It can be enhanced used by implementing in homomorphic algorithm, and by linking it in various applications. It can be implemented by linking many online shopping web application. And also the data contributors trust can be still made strong using various technology.

References

[1] Microsoft Azure Marketplace (2017). [Online]. Available: https:// datamarket.azure.com/home/
[2] Gnip, (2017). [Online]. Available: https://gnip.com
[3] DataSift, (2017).[Online].Available: http://datasift.com/
[4] Datacoup,(2017).[Online].Available:https://datacoup.com/
Citizenme,(2017).[Online].Available:https://www.citizenme.com/
[5] Gallup Poll, (2017). [Online]. Available: http://www.gallup.com/
[6] M. Barbaro, T. Zeller, and S. Hansell, A Face is Exposed for AOL Searcher no. 4417749, New York, NY, USA: New York Times, Aug. 2006.
[7] 2016 TRUSTe/NCSA Consumer Privacy Infographic - US Edition, (2017).
[Online]. Available:https://www.truste.com/resources/privacy-research/ncsa- consumer-privacy-index-us/
[8] K. Ren, W. Lou, K. Kim, and R. Deng, “A novel privacy preserving authentication and access control scheme for pervasive computing environments,” IEEE Trans. Veh. Technol., vol. 55, no. 4, pp. 1373–1384, Jul. 2006.
[9] M. Balazinska, B. Howe, and D. Suciu, “Data markets in the cloud: An opportunity for the database community,” Proc. VLDB Endowment, vol. 4, no. 12, pp. 1482–1485, 2011.
[10] P. Upadhyaya, M. Balazinska, and D. Suciu, “Automatic enforcement of data use policies with datalawyer,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2015, pp. 213–225.