Indian Journal of Science and Technology, *Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/108028, December 2016* *ISSN (Print) : 0974-6846* *ISSN (Online) : 0974-5645*

Secure Retrieval of Sub-Trees over Encrypted Data Stored in Cloud Storage for Prediction

P. Shanthi and A. Umamakeswari

School of Computing, SASTRA University, Thirumalaisamudram, Thanjavur - 613401, Tamil Nadu, India;
shanthip@cse.sastra.edu, aum@cse.sastra.edu

Abstract

Objectives: A tremendous growth in adoption of cloud storage services is observed since 2010. Shifting towards the services however leads to some security concerns like data leakage, unauthorized access and data privacy etc. The main objective is to develop a subtree retrieval method that securely that is further used for prediction without impacting performance**. Methods:** Cryptographic methods are used to secure data. In our approach the dataset is partitioned based on classification and encrypted before uploading to cloud. Retrieval query returns a subtree from the partitions only when the secret key is matched. **Findings:** Our proposed novel approach results in better performance compared to other methods of partitioning. Also securely retrieves subtree over the encrypted data. **Application/Improvements:** This approach can be applied to classify patient electronic health records in cloud storage and query the encrypted data in order to make decisions.

Keywords: Classification, Cloud Storage, Cryptography, Subtree Retrieval

1. Introduction

Cloud Computing provides shared computing resources such as networks, servers, storage facilities and applications as services. Resources are given on demand and no need for provider interaction. It is a pay-per-use model where users pay only for the usage of a resource. Additionally, in[1] Cloud Computing attracts users by promising high availability, fault-tolerance, infinite scalability and reduced upfront investment etc.

Cloud outsourcing[2] is a trend of today because it reduces cost and gives access to hi-tech knowledge. Also migrating towards cloud helps the organizations to focus on core competencies. Development of Cloud Sourcing as a emerging area is perhaps very significant factor. Cryptographic techniques are recently applied to ensure privacy and security of shared data. Various models are designed in this view and are discussed in[3]. Cloud storage has become the standard for storing large volume of data, sharing data easily, and accessing from anywhere. Cryptography techniques play a major role in protecting cloud storages. Different types of cryptography-based storages are designed and proven to be secure. Encryption is used to secure data in semi-trusted storage server. Using single key to encrypt data is insufficient in sharing environment. New encryption method called functional encryption presented in[4] allows getting partial plaintext of stored data and also sharing encrypted data with recipients whose identities are unknown. As many data intensive applications are migrated to cloud, some intermediate datasets are accommodated along with the applications to avoid cost of re-computing them. This may lead to leaking of private information of dataset to malicious users unknowingly. To avoid this problem, conceal all the datasets by encryption before sending to cloud storage. Different levels of encrypting sensitive data are row, column and by page. Column level encryption places sensitive attributes of records altogether in a single page. Searching encrypted sensitive attributes inside a single page incurs less cost compared to other levels of granularity as shown in[5]. Security model for conjunctive keyword search for practical relational database is

**Author for correspondence*

important. Possible vulnerabilities and countermeasures are discussed by in[6]. In[7] vertically (column level) partitioned dataset is distributed to multiple parties where the predicable attribute is with only one party. A decision tree constructed for the distributed dataset preserves the privacy of the data. Decision trees are built using random projections of heterogeneously distributed data to estimate the dot product between two binary vectors. This includes strategies for optimizing message communication but increases computation time[8]. Subtrees of tree set are dealt in different ways. Pruning subtree to discover the structural commonalities between sub-trees and identifying sub-tree semantic resemblances are presented in[9]. This proposed work focuses on retrieving the subtree from the decision tree of encrypted data in order to make a prediction. This paper is organized such that section 2 discusses the various existing works related to the problem, section 3, the contribution towards the problem research area, following that section 4 gives the methodology used to achieve the solution, section 5 throws the conclusion and the scope of future work and finally section 6 concludes with the references used in the progress.

A new public key encryption system presented in realized to be suitable for cloud storage is Functional encryption. Identity based encryption uses public key and the message as input for encryption. Also recipients information is not needed for sharing the message as presented in[10]. Ciphertext policy attribute-based encryption, a kind of Attribute based encryption, data is encrypted by embedding a policy defined over attributes of users and receivers whose attributes matches the policy can get the data was presented in[11]. In this work, the retrieving subtree from the database partitions on matching the secret key of the partitions has been dealt with. Decision tree is traversed from the root to the leaf by following the path that matches the query. When dataset is very large partitioning and processing improves performance. But the choice of partitioning makes physical design automation more complex. In order to predict the value of the class attribute based on the query, classification based partitioning improves performance. In[12], horizontal partitioning is used when searching operation is done on the dataset. TPC benchmark projected in[13] the difficulty of allocating horizontal partitions in a warehouse environment. Various issues of fragmenting as well as allocating to disks are discussed in their paper. Several papers described the vertical partitioning approach and are presented in[14-16]. Their paper studies the partitioning based on the classification of the dataset to improve the performance of predicting the value as well as to securely compute the prediction by preserving the privacy of the data partitions. Database is vertically partitioned into two pieces and a decision tree classifier is built without compromising the privacy of the data. A method in[17] requires a untrusted third party server which uses scalar product protocol to form the classifier. An integrated approach to reduce the effort of expanding the subtree that would be pruned in next phase has been proposed in[18]. Their method calculates a lower bound on the cost subtree to integrate second pruning with the initial building phase. Predictions are expressed as polynomial of bounded degree. A group of machine learning algorithms that includes homomorphic encryption was proposed and this classification on polynomial approximations takes less gradient descent steps to obtain the solutions. In[19] proposed dependency based partition scheme that groups data points of each server. Non-dependent partitions are computed parallely. Mapping dependencies is carried out and query plan is generated, which facilitates in reducing amount of data transferred. In[20], adaptive encryption system is proposed which allows encrypting and querying selective columns

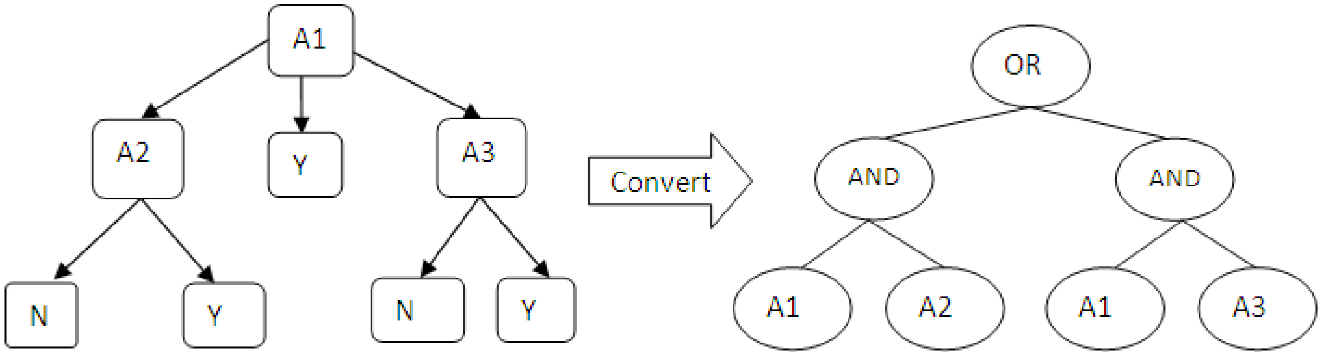

Figure 1. Classification path of the decision tree is converted to Boolean formula as (A1 ∧ A2) ∨ (A1 ∧ A3).

of the dataset. In their system, queries are not decided at design time. A neural security model in[21], used two components for data security. A sensitive data component for storing fragmented data and Counter propagation neural network component for encrypting and decrypting data. They achieved high data confidentiality and also better performance compared to back propagation method.

2. Contribution

Build a private key for the given encrypted dataset that gives the subtree of the encrypted dataset but reveals nothing about data. Decryption key enables the user to get a decision tree for prediction over the encrypted dataset and reveals nothing about data. A trusted authority generates a master secret key. The authority is given the concept tree of the encrypted dataset as input to generate a derived secret key. Any user having secret key is eligible to get the decision tree of the encrypted dataset. The security of the system guarantees that only the secret key holder can retrieve a decision tree of the encrypted dataset.

3. Retrieval Model

This works focuses on partitioning the dataset based on classification. Decision tree is constructed to get the classification for the dataset. The classification is represented as Boolean formula. Number of clauses in the Disjunctive Normal form determines the number of partitions. Dataset is partitioned and is encrypted and the clauses are embedded over the encrypted data. Encrypted data is uploaded to cloud storage. Secret key to predict value is derived from the master key. On giving the generated key to encrypted dataset, the subtree of the dataset needed for the prediction is retrieved and finally returns the prediction as probability percentage.

4. Tree Construction

Decision tree (DT) is constructed for the given dataset using ID3 algorithm, a top-down search through the possible branches with no backtracking. ID3 uses two metrics, Entropy and Information Gain to construct a decision tree. Entropy is a measure of disorder or impurity. Entropy of the output values of a set of training instances are found and is the average number of bits needed to encode an output value. The entropy changes when a node in a decision tree is chosen to partition the training instances into smaller subsets. Information gain is a measure of this change in entropy.

Algorithm 1 Decision Tree (FV,PA,D)

```
Requirement: set of feature vectors (FV), Predictable Attribute (PA), a dataset
Result: a decision tree

    • If FV is empty then return default
    • Else if all tuples in D have the same class then return the class
    • Else
    • Determine the attribute A in FV that best
    •           classifies the tuples in D
    • Assign tree as new decision tree with root
    •           as best
    • For each value $s_i$ of best do
    • $D_j$□ (tuples of D with best = $s_i$)
    • subtree□ DecisionTree($D_j$,FV-best)
    • Add a branch to tree with label $s_i$ and subtree subtree
    • Return tree
```

4.1 Converting DT to Boolean Formula

In decision tree, leaves represent predictable attributes and branches represent conjunctions of features that lead to those predictable attributes. DT is converted to Disjunctive Normal Form (DNF). Each clause in the DNF gives the classification over the dataset. Dataset is partitioned based on the description of classification clause and are encrypted by embedding them. Figure 1 shows the conversion of DT to AND-OR tree and dataset is partitioned vertically based on the number of clauses. Processing partitioned data incurs more cost. Partitioning data using an approach that reduces processing is a challenge. Table 1 shows the number of partitions for the given dataset. Partition increases as the number of tuples increases. The number of partitions can be reduced by selecting best splitting criterion. Thus optimizing tree path improves performance.

4.2 Subtree Retrieval from Encrypted Dataset Stored in Cloud Storage

Retrieving subtree from the encrypted dataset involves four algorithms:

Setup(λ): System authority takes security parameter as input, and generates master key MKY and public parameters as output.

Keygen (MKY,FV): Uses master key MKY and Feature_vector set (FV) for retrieving the subtree of decision tree as input and produces a secret key pv[f] specific to the retrieval of partitioned dataset. Precisely, given correct secret key, user is allowed to access only the attribute set authorized by the key.

Encyption (MSK, CA): Takes public parameters, master key and dataset classification in Boolean form as input and generates the encrypted dataset. Dataset is partitioned based on the classification (CA) and encrypted in functional encryption. Function f embeds the attributes of each classification of the dataset. Functionally encrypted dataset is outsourced to cloud server.

Decryption (SK,CT): Requires secret key (SK) and ciphertext (CT) as input. First it produces a function for the decryption of dataset. Function f, on authorizing the key, retrieves a subset of dataset that matches the given key as output. User is given the capability to predict the value of the subtree of encrypted dataset only when the key is correct. On matching the key with the attributes, a set of partitioned data is retrieved and manipulated to construct the subtree that gives the prediction value.

Table 1. Number of partitions for different sample datasets

Name of Dataset	Number of Attributes	Number of Partitions	Number of tuples
Golf dataset	4	2	14
Iris dataset	4	>3	150
Labor negotiation dataset	16	1	40
Weighing dataset	6	>5	500
Ripley dataset	2	1	250

4.3 Security of the System

This model is based on functional encryption, and hence the same security level is applicable to this also. An adversary having secret keys vp[f1],…,vp[fn] for performing operations over the encrypted partition, can learn nothing about the decryption of other attributes as this encryption method does not give rights to the partitions other than what is revealed by the keys at his disposal.

Let the adversary define a test T for d=0, 1 as follows:

- Setup: run(ppm.mky)□setup(1^λ) and give ppm to adversary.
- Query: Adversary submits queires q_k in Q for k=1,2,… and is secret key pv_k by keygen process.
- Challenge: Adversary submits partitions $p_i \in X$ is given encryption of m_d.
- Adversary continues to issue key queries as before that outputs a bit in d.

The probability is given by probability of event getting output 0 minus the probability of getting output 1.

5. Conclusion and Future Work

A novel approach for predicting value from the encrypted data stored in cloud storage is proposed. Classification based partitioning preserves privacy by revealing only the partitioning that matches the query and also improves performance by retrieving the subtree that is required to predict the class value. This method is exercised for dataset that is large having tuples approximately 100,000 and 100 attributes. But when dataset is scaled to million or billion, scanning data and construction of tree will take more time. Hence this work can be extended to comply with very large set using tree optimization method.

6. References

1. Services in the cloud computing era: A survey. Date Accessed: 18/10/2010: Available from: http://ieeexplore.ieee.org/document/5666772/.
2. Lacity MC, Khan SA, Willcocks LP. A review of the IT outsourcing literature: Insights for practice. The Journal of Strategic Information Systems. 2009 Sep; 18(3):130-46.
3. Yu S, Wang C, Ren K, Lou W. Achieving secure, scalable, and fine-grained data access control in cloud computing. INFOCOM'10 Proceedings of the 29th conference on Information communications. 2010; p. 534-42.
4. Boneh D, Sahai A, Waters B. Functional encryption: a new vision for public-key cryptography. Communications of the ACM. 2012 Nov; 55(11):56-64.
5. Canim M, Kantarcioglu M, Inan A. Springer Berlin Heidelberg: Query optimization in encrypted relational databases by vertical schema partitioning. 2009 Aug; p. 1-16.
6. Byun JW, Lee DH. On a security model of conjunctive keyword search over encrypted relational database. Journal of Systems and Software. 2011 Aug; 84(8):1364-72.

7. Vaidya J, Clifton C, Kantarcioglu M, Patterson AS. Privacy-preserving decision trees over vertically partitioned data. ACM Transactions on Knowledge Discovery from Data (TKDD). 2008 Oct; 2(3):1-27.
8. Communication efficient construction of decision trees over heterogeneously distributed data. Date Accessed: 1/11/2004. Available from: http://ieeexplore.ieee.org/document/1410268/.
9. Tekli J, Chbeir R. A novel XML document structure comparison framework based-on sub-tree commonalities and label semantics. Web Semantics: Science, Services and Agents on the World Wide Web. 2012 Mar; 11:14-40.
10. Sahai A, Waters B. Fuzzy identity-based encryption. EUROCRYPT'05 Proceedings of the 24th annual international conference on Theory and Applications of Cryptographic Techniques. 2005; p. 457-73.
11. Waters B. Springer Berlin Heidelberg: Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization. 2011 Mar; p. 53-70.
12. Zeller B, Kemper A. Experience report: exploiting advanced database optimization features for Large-Scale SAP R/3 installations. VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases. 2002; p. 894-905.
13. TPC Benchmark. Date Accessed: 05/1999: Available from: http://www.tpc.org/reports/status/bs-1999-05.asp.
14. Cornell DW, Yu PS. An effective approach to vertical partitioning for physical design of relational databases. IEEE Transactions on Software Engineering. 1990 Feb; 16(2):248-58.
15. Navathe SB, Ra M. Vertical partitioning for database design: a graphical algorithm. SIGMOD '89 Proceedings of the 1989 ACM SIGMOD international conference on Management of data. 1989 Jun; p. 440-50.
16. Papadomanolakis S, Ailamaki A. Autopart. Automating schema design for large scientific databases using data partitioning. Proceedings 16th IEEE International Conference on Scientific and Statistical Database Management. 2004 Jun; p. 1-10.
17. Du W, Zhan Z. Building decision tree classifier on private data. Proceedings of the IEEE international conference on Privacy, security and data mining. 2002; 14:1-8.
18. Rastogi R, Shim K. PUBLIC: A decision tree classifier that integrates building and pruning. Data Mining and Knowledge Discovery. 2000 Oct; 4(4):315-44.
19. ML confidential: Machine learning on encrypted data. Date Accessed: 26/12/2012: Available from: https://eprint.iacr.org/2012/323.
20. Wagh JC, Mhatre S. Implementing Encrypted Database for Concurrent Access in Cloud environment. Indian Journal of Science and Technology. 2016 Apr; 9(13):1-8.
21. Jegadeeswari S, Dinadayalan P, Gnanambigai N. Neural based Security Approach for Cloud databases using Counter Propagation. Indian Journal of Science and Technology. 2016 Apr; 9(16):1-10.