Smart Search over Enciphered Data for Cloud Computing

Sarvesh Kher[1] Mayuresh Patil[2] Gaurav Mahendrakar[3] Gauri Kavitkar[4] Nikhita Nerkar[5]
[1,2,3,4]Student [5]Professor
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]RMD College of Engg, Pune, Maharashtra, India

Abstract— With the increasing adoption of cloud computing, a growing number of users outsource their datasets to cloud. Now a day's there is large amount of data on the cloud. Most existing symmetric searchable encryption schemes aim at allowing a user to outsource her encrypted data to a cloud server and delegate the latter to search on her behalf. The concept of searchable encryption provides a promising direction in solving the privacy problem when outsourcing data to the cloud. It allow users to store their data in encryption from at an untrusted server, and then delegate the server to search on their behalf by issuing a private key and encrypted search index.

Key words: Searchable Encryption, Data Outsourcing, Cloud Computing, TF-IDF Algorithm, Search Index

I. INTRODUCTION

The concept of searchable encryption provides a promisingdirection in solving the privacy problem when outsourcingdata to the cloud. Such schemes allow users to store their datain encrypted form at an untrusted server, and then delegatethe server to search on their behalf by issuing a trapdoor(i.e. encrypted keyword).[1,7] Sometimes, a user wants to verify whether the data it wants to access is the same data that was initially uploaded on the cloud. For this, a Scheme needs to be publicly verifiable [3].

Theindividual can remotely store the data on the cloud server,namely data outsourcing, and then make the cloud data openfor public access through the cloud server. This represents a more scalable, low-cost and stable way for public data accessbecause of the scalability and high efficiency of cloud servers, and therefore is favorable to small enterprises.

Outsourcing of data or time-consuming computational workloads on the cloud solves the problems of maintenance, reduces the needless repetition of data information, which decrease the burden on individuals or enterprises/institutions. Due to the private nature of personal data, there is an inherent need for a user to selectively share the data with different recipients.[11]

Suppose user A wants to share his private key with some user B, user A will encrypt his private key using public key of user B and send it to user B so no other person except user B can decrypt that message. [9] For searching TF-IDF algorithm technique will provide excellent time efficiency. Server will maintain clusters for frequently searched keywords.To design a content-based search theme and build linguistics search more practical and context-aware could be a tough challenge. Several systems area unit projected to form encrypted knowledge searchable supported keywords. However, keyword-based search schemes ignore the linguistics illustration info of user's retrieval, and can't fully meet with users search intention. Authorization information along with index file of each document helps to detect malicious users and prevent them from accessing user's private data [1, 2, and 6].

To understand that the system is simply able to search over encrypted data however not in cloud computing. Therefore there's a desire to develop a system which may linguistics search over encrypted data for cloud computing. Also, the system may be developed for knowledge storing and retrieving from the cloud with economical key management and sharing techniques

II. LITERATURE REVIEW

A. Multi-User Schemes

Curtmola et al.[1]proposed the concept of multi-user searchable encryptionschemes, where a user can authorize multiple other users tosearch the encrypted data. However, the proposed primitivedoes not take into account the fact that the same user may alsobe authorized to search other users' data and the correspondingsecurity issues. As a result, the primitive from [1] offers asolution for a much more simplified problem than ours, andit seems not trivial to construct a scalable solution for ourproblem based on their scheme.

Divide tasks into the three entities into Group Manager, Opening Manager and Revocation Manager, [3] toincrease the privacy of the user by dividing the work of the group manager into three entities. Group Manager can only create group and add members in group but does not possesses power to open any signature. The Open Manager possesses a special key which can be used only to open a signed message.

The main idea behind our scheme is that the secret key of the group [8] is split into two parts by GM, one part is given to the user as his group membership secret key, and the other is given to SEM. Neither the group member nor SEM can sign a message without the other's help. To revoke the membership of a group member, GM need only ask SEM not to provide the group member partial signatures any more.

The paper [5] propose a systematic solution, which refers to as QDMiner, to automatically mine query facets by extracting and grouping frequent lists from free text, HTML tags, and repeat regions within top search results. Experimental results show that a large number of lists do exist and useful query facets can be mined by QDMiner. They further analyze the problem of list duplication and find better query facets can be mined by modeling fine-grained similarities between lists and penalizing the duplicated lists.

In [10] paper, the author considers objects that are tagged with keywords and are embedded in a vector space. For these datasets, they study queries that ask for the tightest groups of points satisfying a given set of keywords. It proposes a novel method called ProMiSH (Projection and Multi-Scale Hashing) that uses random projection and hash-based index structures and achieves high scalability and

speedup. Also, they present an exact and an approximate version of the algorithm. The experimental results on real and synthetic datasets show that ProMiSH has up to 60 times of speedup over state-of-the-art tree-based techniques.

Long et al. [2] proposed algorithms to retrieve a group of spatial web objects such that the group's keywords cover the query's keywords and the objects in the group are nearest to the query location and have the lowest inter-object distances. Other related queries include aggregate nearest keyword search in spatial databases. First, existing works mainly focus on the type of queries where the coordinates of query points are known [5], [2].

Subhra Mishra and TilakRajanSahoo[7] The scheme implemented by us provides these features. The use of elliptic curve cryptography increases the security the scheme by providing desired security level that is achieved by significantly smaller keys in elliptic curve system than in its counterpart- RSA system. Another significant advantage being in general, the algorithms used for encryption and decryption in ECC schemes are faster and can be run on machines that are less efficient.

The secret key of the group is split into two parts by GM, one part is given to the user as his group membership secret key, and the other is given to SEM. Neither the group member nor SEM can sign a message without the other's help. To revoke the membership of a group member, GM need only ask SEM not to provide the group member partial signatures any more.

III. PROPOSED SYSTEM

In proposed system users doesn't need to share private key with server, so server cannot decrypt user's data and data confidentiality remains high. User can self-generate his private key after choosing a public key. User can encrypt each different document using different public keys by this way user can provide more security to each of his text document. So the new proposed scheme provides higher level of security in searching valuable information which is to be shared by multiple users. Authorization information along with index file of each document helps to detect malicious users and prevent them from accessing user's private data.

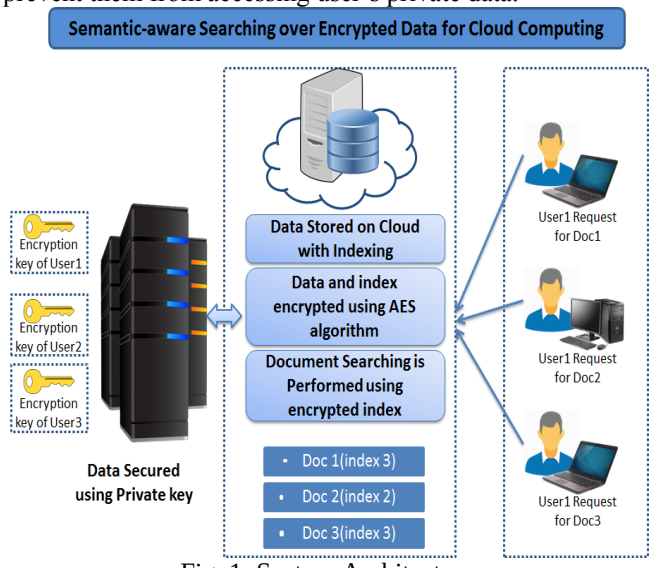

Fig. 1: System Architecture

1) Store documents on cloud

Store documents on cloud with mapping the index. All Documents are encrypted along with the search index. Search index is given to each document for fast searching.

2) Key Exchange with User

Generate keys for single user's document, a single key having multiple authorization codes will be used for decryption purpose. Each user will have to sign in to our system; its role and department will be extracted from login information. Department and role will be used for extracting user's trapdoor key from the key. Once the trapdoor key is extracted, it will used for searching andlocating the user document.

3) Key Management (RSA Algorithm)

Key management is done with the help of RSA Algorithm. Apply RSA Algorithm for public private key. Data will encrypt using public private key.

4) Multi-party searchable encryption

When user gets request for searching document. That time user sends search index according to the cloud server. Once the data will searched then data will decrypt using AES algorithm.

5) Content Based Filtering using TF-IDF

a) Searching data is divided into several attributes example Item U may be having attributes A1,A2,A3,A4...An.
b) We have several items in the database may be U1, U2, U3...UN.
c) Each item's attributes are compared to rest of the items in the database and a cumulative score is calculated based on their similarities.

Hence an algorithm is used to match these attributes example if U1 has A1 as A,B,C and U2 has A1 as B,C then their matching score would be U1(A1)intersection U2(A1)/#of U1(A1).

IV. ALGORITHM USED

The overview of the different algorithm used by the researcher in the previous paper is given below.

1) Encryption/Decryption Using AES Algorithm
2) TF-IDF Algorithm
3) SHA-1(Secure Hash Algorithm 1)
4) ECC(Elliptic-curve cryptography)

1) Encryption/Decryption Using AES Algorithm:[14]

– AES is based on a design principle known as a substitution-permutation network and is fast in both software and hardware. Unlike its predecessor DES, AES does not use a Feistel network. AES is a variant of Rijndael which has a fixed block size of 128 bits, and a key size of 128 bits. By contrast, the Rijndael specification per se is specified with block and key sizes that may be any multiple of 32 bits, both with a minimum of 128 and a maximum of 256 bits.
– AES operates on a 4×4 column-major order matrix of bytes, termed the state, although some versions of Rijndael have a larger block size and have additional columns in the state. Most AES calculations are done in a special finite field.
– The key size used for an AES cipher specifies the number of repetitions of transformation rounds that convert the input, called the plaintext, into the final output, called the

ciphertext. The number of cycles of repetition is as follows:

- 10 cycles of repetition for 128-bit keys.
- Each round consists of several processing steps, each containing four similar but different stages, including one that depends on the encryption key itself. A set of reverse rounds are applied to transform ciphertext back into the original plaintext using the same encryption key.

2) TF-IDF Algorithm:[14]

- TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term.
- The TF*IDF algorithm is used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document. More importantly, it checks how relevant the keyword is throughout the web, which is referred to as corpus.

For a term t in document d, the weight Wt,d of term t in document d is given by:

Wt,d = TFt,d log (N/DFt)

Where:

- TFt,d is the number of occurrences of t in document d.
- DFt is the number of documents containing the term t.
- N is the total number of documents in the corpus.

3) Secure Hashing Algorithm-1: [14]

- In the domain of cryptography, Secure Hash Algorithm 1 is a cryptographic hash function designed by the United States National Security Agency and is a U.S. Federal Information Processing Standard published by the United States NIST.

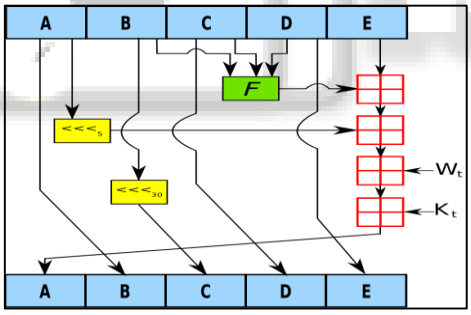

Fig.2: Pin Diagram of SHA-1[15]

- SHA-1 produces a 160-bit (20-byte) hash value known as a message digest. A SHA-1 hash value is typically rendered as a hexadecimal number, 40 digits long.

The above figure describes single iteration within the SHA-1 compression function: A, B, C, D and E are 32-bit words of the state; F is a nonlinear function that varies; n denotes a left bit rotation by n places; n varies for each operation; Wt is the expanded message word of round t; Kt is the round constant of round t; denotes addition modulo 232.

4) ECC Algorithm:

- Elliptic-curve cryptography (ECC) is an approach to public-key cryptography based on the algebraic structure of elliptic curves over finite fields. Elliptic curves can be used for encryption by combining the key agreement with a symmetric encryption scheme.
- In today's world ECC algorithm is used in case of key exchanges by certificate authority (CA) to share the public key certificates with end users. Elliptic Curve Cryptography is a secure and more efficient encryption algorithm than RSA.
- The security of ECC algorithm depends on its ability to compute a new point on the curve given the product points and encrypt this point as information to be exchanged between the end users.
- The ECC system is based on the concepts of Elliptic Curves. To analyze the time taken by an algorithm researches have introduced polynomial time algorithms and exponential time algorithms. Algorithms with smaller computation can be evaluated with polynomial time algorithms and complex computations can be evaluated with exponential time algorithms.

The fig. 2 shows a simple elliptic curve.

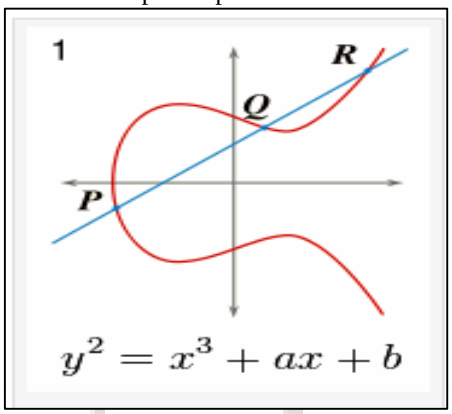

Fig. 3: Graph of Elliptic Curve [16]

The equation of an elliptic curve is given as,

y2 = x3 + ax + b

a) Key generation:

Key generation is an important part where an algorithm should generate both public key and private key. The sender will be encrypting the message with receiver's public key and the receiver will decrypt its private key. Now, select a number, d within the range of n. Generate the public key using the following equation,

$$Q = d * P$$

Where d = the random number selected within the range of (1to n-1). P is the point on the curve, Q is the public key and d is the private key.

b) Encryption

Use the following equation to get back the original message 'm' that was sent.

M = C2 - d * C1

c) Decryption

Use the following equation to get back the original message 'm' that was sent.

M = C2 - d * C

M is the original meaasage that was sent.

V. Experimental Results and Discussion

In Signing table, 19 random samples are used for Encryption, Decryption, Key Generation and Signing process.We have formulated a new primitive, smart search over Enciphered data, for enabling users to selectively authorize each other to search in their encrypted data which gives a better improvement than existing system.

Encryption, Decryption, Key Generation and Signing			
Generation(ms)	Encryption(ms)	Decryption(ms)	Signing(ms)
7.870686	6.541971	5.177336	7.417747
5.54954	5.15681	5.01826	4.030618
4.279666	4.315587	5.890272	2.001967
4.303956	4.38777	5.695617	1.723498
5.224545	5.876588	4.782553	1.284926
5.080864	5.003207	4.563609	1.164507
4.980287	5.080522	5.554672	1.114218
4.293351	4.205773	5.463332	1.182639
4.161985	4.314561	4.862947	1.051272
4.727133	4.775369	4.35527	0.9613
4.678897	4.594056	4.366559	0.941116
4.686423	4.613555	4.43498	1.495318
4.164721	4.452426	5.040154	0.902117
4.077143	4.969682	5.010049	0.915458
5.108232	4.723369	4.55232	1.023562
4.74458	4.676502	4.415138	0.935301
4.713791	4.04738	4.875946	0.900064
4.076117	4.353902	4.913235	0.890143
4.103143	4.140432	4.926919	0.882617
in ms	4.780266316	4.748919053	4.942061474	1.622020421

Fig. 4: Encryption, Decryption and Signings

Algorithm Comparision		
Sr.No	Algorithm	Time(ms)
Encrypt Data	DES	24
	AES	17
	ECC	7
Decryption	DES	28
	AES	20
	ECC	9

Fig. 5: Algorithm Comparision (Table)

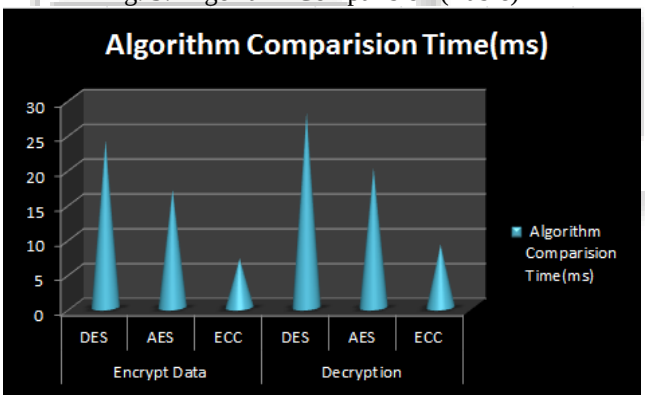

Fig. 6: Graphical Representation of Algorithm

VI. CONCLUSION

The system uses one cloud server for encrypted retrieval of data and makes contributions both on search accuracyand efficiency. To improve accuracy, we extend the concept hierarchy to expand the search conditions.Authorization information along with index file of each document helps to detect malicious users and prevent them from accessing user's private data.Experiments on real world dataset illustrate that our scheme is efficient.

Many systems are proposed to make encrypted data searchable based on keywords. However, keyword-basedsearch schemes ignore the semantic representation information of user's retrieval, and cannot completely meet with usersearch intention.

REFERENCES

[1] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchablesymmetric encryption: Improved definitions and efficient constructions," in Proc. 13th ACM Conf. Comput. Commun. Security, 2006, pp. 79–88.
[2] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, "Collective spatialkeyword querying," in Proc. ACM SIGMOD Int. Conf. Manage.Data, 2011, pp. 373–384.
[3] Zhicheng Dou, Member, IEEE, Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song, "Automatically Mining Facets for Queries from Their Search Results", IEEE Transactions on Knowledge And Data Engineering, Vol. 28, No. 2, February 2016.
[4] Hongwei Li, Member, IEEE, Yi Yang, Tom H. Luan, Xiaohui Liang, Liang Zhou, and Xuemin (Sherman) Shen, "Enabling Fine-grained Multi-keyword Search Supporting Classified Sub-dictionaries over Encrypted Cloud Data", IEEE Transactions On Dependable And Secure Computing, Vol. 13, No. 3, May/June 2016.
[5] C. Long, R. C.-W. Wong, K. Wang, and A. W.-C. Fu, "Collectivespatial keyword queries: A distance owner-driven approach," inProc. ACM SIGMOD Int. Conf. Manage. Data, 2013, pp. 689–700.
[6] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preservingpublic auditing for data storage security in cloud computing,"inProc. of IEEE INFOCOM 2010, CA, USA, Mar. 2010, pp. 525–533.
[7] Pushkar Zagade, Shruti Yadav, Aishwarya Shah, Ravindra Bachate " Group User Revocation and Integrity Auditing of Shared Data in Cloud Environment" International Journal of Computer Applications (0975 – 8887) Volume 128 – No.12, October 2015.
[8] Subhra Mishra and Tilak Rajan Sahoo" A Survey on Group Signature Schemes" Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela.
[9] Tao Jiang, Xiaofeng Chen, and Jianfeng Ma" Public Integrity Auditing for Shared Dynamic Cloud Data with Group User Revocation" 2015 IEEE.
[10] He Ge "An Effective Method to Implement Group Signature with Revocation".
[11] S. Cui, X. Cheng and C. W. Chan, "Practical group signatures from RSA," 20th International Conference on Advanced Information Networking and Applications - Volume 1 (AINA'06), Vienna, 2006
[12] Vishwakarma Singh, Bo Zong, and Ambuj K. Singh, "Nearest Keyword Set Search in Multi-Dimensional Datasets", IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 3, March 2016.
[13] Zhangjie Fu Lili Xia Xingming Sun Alex X. Liu Guowu Xie, "Semantic-aware Searching over Encrypted Data for Cloud Computing", IEEE Transactions on Information Forensics and Security, 2018.
[14] Sarvesh Kher,Mayuresh Patil,Gaurav Mahendrakar,Gauri Kavitkar and Nikhita Nerkar "Smart Search over Enciphered Data for Cloud Computing: A Survey" IJRASET.
[15] https://en.wikipedia.org/wiki/SHA-1
[16] https://en.wikipedia.org/wiki/Elliptic-curve_cryptography

