ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 15, March 2016

MULTICLOUD STORED ENCRYPTED BIG DATA SECURED BY HONEY WORDS

P.Shanmugavalli
PG Student, CSE Department
Anna University Regional Office, Madurai
Madurai, India
Shanmugavalli79@yahoo.in

Dr.S.Srinivasan M.Tech,Ph.D
Professor and Head, CSE Department
Anna University Regional Office, Madurai
Madurai, India
sriniss@yahoo.com

Abstract— **The Big data has a great impact on business as it decides the future business. But it faces many challenges such as availability and security. Future products and services could be developed only on the data that is available with high integrity. In E-commerce a large amount of data is produced by customers. This work proposes a multicloud environment to securely store big data. This cloud storage provides robustness, availability and avoids single point failure. The important criteria for big data are its continuous availability for analytical processing. The data is split and stored in different cloud storages. Some parts of the data are encrypted because encrypting the whole big data is a tedious and a very costly process. A part of a large data will not be fruitful for the adversaries. The storage path of the data part is stored in a path file which is trapdoor secured. The tenant (merchant) after authorizing with the data owner will get the path file. Then the tenant will download all data parts and merge and perform analytical process over that data. This provides the information regarding his product's sale. This in turn helps the merchant to improve his service and find his success rate. If an unauthorized (competitor) access is made an alarm triggers at the data owner's site. Hence a breach process is known to the data owner. This in turn protects the valuable data. It is also ensured that tenant is not allowed to change any data parts.**

Keywords—Big Data, Multicloud, trapdoor, honeyword.

I. INTRODUCTION

The big data itself contains a term related to size which is an important characteristic of big data. The amount of data that is being created and stored is so large and it just keeps growing. Big data refers to large dataset which is complex that traditional data processing applications can't process. Big data usually includes data sets that commonly used software tools are not applicable to capture, curate, manage, and process data within a tolerable elapsed time. Cloud computing is computing based on the internet. In the past, the applications or programs are downloaded on a computer or server user's destination. As the data is increasing day by day normal storage of data in physical computer becomes impossible. It allows people to access the same kinds of applications through the internet. The advantage of cloud computing is its low cost and accessibility of data. Ensuring the security of cloud computing is a major factor in the cloud computing environment, as users store sensitive information in cloud storage but these providers may be untrusted. The protection of user's privacy is the biggest challenge for big data from a security point of view. Big data contains personal identifiable information stored in cloud and therefore privacy of users is a huge concern. Because of the importance of big amount of data stored, breaches affecting big data can have more devastating consequences than the normal data breaches .This is because a big data security breach will affect a larger number of people, not only from a reputational point of view, but with enormous legal repercussions.

II. RELATED WORK

An important role in today's information systems is played by data centres which always perform computations that are complex and retrieve large amount of datasets from data centres. In a distributed environment, an application make use of different datasets located in various data centres and therefore face some challenges such as data security, privacy protection and authentication. Ensuring the cloud computing security is a major factor in the cloud computing environment, as the stored information is sensitive. A single cloud provider is predicted to be less popular with customers because they have the risks of service availability failure and the possibility of malicious insiders in the single cloud. Hence M.A.Alzain et.al [1] proposed multi-clouds due to its ability to reduce security risks that affect the cloud computing user .

Chang- Ji Wang et.al [2] proposed that a PHR data can be stored in a cloud and can be used by different users such as doctor ,nurse ,friends and family. In order to store the data securely, the data is portioned into two domains, public and personal. Two types of encryption are used for these domains . They are ABE for public domain and anonymous multi-receiver IBE for personal domain. But encrypting all the data is a tedious process as the data keeps on increasing. Kan Yang et.al [7] proposed data access control scheme for multi-authority cloud storage systems .A revocable multi-authority CP-ABE scheme was designed for data access control. The drawback of policy updating is once data owner outsource the data into cloud, no copy is stored in data owner's systems. In order to change access polices the whole process has to repeated such as data retrieval and encryption which needs to be send back to cloud that incurs high communication overhead and heavy computation burden. S. Yakoubov et.al [9] discuss about a computation model for big data analytics in the cloud and several cryptographic techniques that can be used to secure these analytics in a variety of settings were proposed. Three cryptographic techniques – homomorphic encryption, verifiable computation, and multi-party computation. If the data increases exponentially, encrypting cost also increases exponentially. Ning Cao et.al [6] proposed multi keyword ranked search over encrypted cloud data (MRSE). Here a "coordinate matching technique is used for searching. Likewise "inner product similarity" is used for similarity measure. Intrusion detection systems (IDSs) lack the following such as information overload, unknown attacks, false positives and false negatives. So Zhi-Hong Tian et.al [10] proposed the design of AAIDHP (an architecture for intrusion detection using honey pot). I. Erguler [5] proposed that, for each user account, a legitimate password and several honey words are stored randomly in order to sense impersonation. Logging in with a honeyword will trigger an alarm indicating the administrator about a password file breach. The author introduces a simple and effective solution for the detection of password file disclosure events at the expense of storage cost. H. Ulusoy et.al [8] proposed that honey words will be used to detect any unauthorized access of data stored in Map Reduce systems. It is discussed how Map Reduce framework can be enhanced with honey pot traps to send alarms to the data controller if a honey data is accessed without authorization. But the drawback of this system is real data and honey data leads to twice the size of real data which requires more space. Also single point failure might end up in crash of the entire system. Cheng Hongbing et.al [3] proposed an approach which divides big data into sequenced parts and stores them in multiple cloud storage service providers. Instead of securing the whole big data, the proposed scheme protects the mapping function of the various data elements to each provider using a trapdoor function. Generally, individual data part is not so significant for adversaries to learn the privacy of big data.

III.PROPOSED SYSTEM

In this work, at Data Owner's site the large data set is first split into small files of equal size. Then these files are uploaded into the cloud in an orderly fashion of equal numbers. This storage path is stored in a path file which is trapdoor protected. When tenant's wants to access these data they get the path file, merge all the files based on the storage location of the files in path file. Then they perform the analytical process on the downloaded data. If an unauthorized access is made to the cloud or path file, then an alarm will be triggered at the data owner's site. Fig 1 shows the architecture of the proposed system.

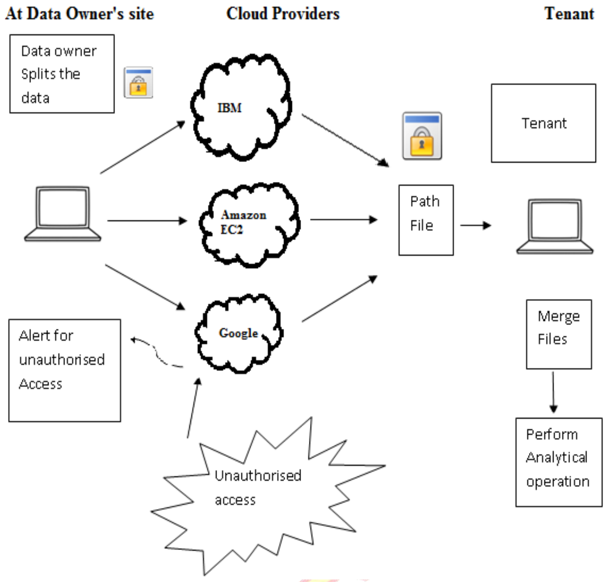

Figure 1: Architecture diagram

The modules in this system are:

A. *Splitting Data*

In order to process the large data, initially the randomly generated data set is divided into small chunks. These data parts are of equal size such as 125kb.This split up makes the file transfer easier. Certain data parts are encrypted using symmetric encryption.

B. *Storing the data in different cloud providers*

The splitted parts will be mapped to different cloud providers such as Google, IBM, and Amazon EC2.This means only certain part of the whole data will be present in a single provider.

C. *Storing the path of data in the path file*

The mapping of the data part to the cloud providers is stored in a path file. Equation (1) provides

$$Mapping_{\,Storage\ Path} = \{P1.m1,\ P2.m2,\ Pn.mn\} \qquad (1)$$

where, P denotes the storage provider, and m denotes the part Of the data stored and n denotes the number of parts. This path file is protected by means of a trapdoor function which could be securely shared to the tenants.

A trapdoor function is a function that is easy to compute in forward direction, but difficult to compute in the reverse direction (finding its inverse) without special function, called the "trapdoor". Trapdoor functions are mainly used in cryptography. In mathematical terms, if f is a trapdoor function then there exists a secret function y, such that given f(x) and y it is easy to compute x. A map is represented by (2)

$$: G1 \times G1 \rightarrow G2 \qquad (2)$$

is called a bilinear pairing if, for all x, y ε G1 and all a, b ε Z , we have (3)

$$(x^{a}, y^{b}) = \quad (x, y)^{ab} \qquad (3)$$

The Bilinear-Diffie-Hellman problem (BDH) for a bilinear map is given by (4)

$$: G1 \times G1 \rightarrow G2 \qquad (4)$$

such that |G1 | = |G2 | = q is prime which is defined as follows: given g, g^{a}, g^{b}, g^{c} →G1 , compute $(g, g)^{abc}$, where g is a generator and a, b, c ε Z .

An algorithm A is said to solve the BDH problem with advantage ε if (5) is satisfied

$$\Pr[A(g, g^{a}, g^{b}, g^{c}) = \quad (g, g)^{abc}] \geq \varepsilon \qquad (5)$$

where the probability is over the random choice of a,b,c,g, and the random bits of A.

D. *Merging the data by tenant*

When the tenant wants to perform an operation, he will get the path file that is trapdoor protected by means of Identity based Encryption .Then with the secret information he will be able to see the distribution of data parts in cloud. He will download all those files and merge it in his site.

The procedure of establishing communication key between Data Owner say A and tenant say B,

Data Owner *A computes parameter Y_A:*

Choose $X_A < q$

Compute (6)

$$Y_A = \eta^{X}{}_{A} \bmod q \qquad (6)$$

Tenant B computes parameter Y_B :

Choose $X_B < q$

Compute (7)

$$Y_B = \eta^{X}{}_{B} \bmod q \qquad (7)$$

Data Owner A encrypts Y_A ,IDA and IDB using IBE algorithm and then send to B:

Encrypt (Y_A ,IDA and IDB)→B

Tenant B encrypts Y_B ,IDB and IDA using IBE algorithm and then send to B:

Encrypt (Y_B ,IDB and IDA)→A

Data Owner A decrypts message and compute (8)

$$K_1 = (Y_B)^{X}{}_{A} \bmod q \qquad (8)$$

Tenant B decrypts message and compute (9)

$$K_2 = (Y_A)^{X}{}_{B} \bmod q \qquad (9)$$

From the above algorithm, (10) concludes that

$$K_1 = (Y_B)^{X}{}_{A} \bmod q = \eta^{X}{}_{A}{}^{X}{}_{B} \bmod q$$
$$= (Y_A)^{X}{}_{B} \bmod q = K_2 \qquad (10)$$

E. *Performing analytical operation*

Now the downloaded file is ready for operation. Any form of analytical operations could be performed in the data present. In this project a simple search processing using map reduce is done. But only certain parts are encrypted so a multi-keyword ranked search over encrypted data based on hierarchical clustering index (MRSE-HCI) [4] to maintain the close relationship between different plain data parts along with encrypted parts in order to enhance the search efficiency is implemented.

F. Alarm for unauthorized access

When an unauthorized person try to access the path file or cloud, then honey word will trigger an alarm at the data owners site. With this we could know that a breach has occurred**.**

For each tenant account, the correct password is stored along with several impersonated honey words. If honey words are generated properly, a cyber-attacker who steals a file which contains hashed passwords cannot be sure if it is the real password or a honeyword for any account. Moreover, logging with a honeyword to login will trigger an alarm indicating the administrator about a password file breach. Instead of increasing the storage requirement, a simple and effective solution to the detection of unauthorized access.

In Chaffing-with-a-password-model the generator algorithm takes the password from the user and relying on a probabilistic model of real passwords it produces the honey words. For instance, mice3blind is decomposed as 4-letters + 1-digit + 5-letters →L4+D1+L5 and replaced with the same composition like gold5rings.

IV.RESULTS

In E-Commerce, large amount of data are generated by the customers. Usually the Sellers or data owners store these data in a cloud due to its volume. These data are processed in order to know the customer preferences for a particular product. In the proposed system we take an imaginary dataset as seller information of various soaps. From that data we would know the highest selling soap brand. If the soap producers (tenants) want to know about their sales they could get access from the seller and perform analytical operation. Big data paves way for the future development of products and services.

In the proposed scheme of cloud storage, the big data of the data owner is divided into a sequence of n parts, where each part can be denoted by *part i* where i range from 1 to n and they will be stored at m different storage providers. Evidently, n is always greater than m, these m storage providers belong to different organizations, such as Google, Amazon, IBM and Yahoo. Each data part stored in these cloud storage providers will be allocated to some storage locations that belongs to the storage provider, so, when big data of a data owner is stored, it will form a unique storage path for the big data given as $Mapping_{\ Storage\ Path}= \{P1.m1,\ P2.m2,\ ....,\ Pn.mn\}$; where, P denotes the storage provider, and m denotes the storage location.

The proposed scheme store the sequenced data parts of big data into cloud storage providers, it also stores redundant data on different cloud storage providers in order to enhance the availability of the big data. The data owner will keep the storage path secretly, and will only share the path mapping with the authorized tenants who genuinely register with the data owner.

Analytical operations could be done on these data similar to the one stored in single cloud. Hadoop framework for mapreduce is used for performing analytical operation. All the data parts are gathered at tenant site and mapreduce is done. Before mapreduce is been done, the encrypted part should be decrypted. For finding that encrypted part we use MRSE-HCI which in turn can be used to decrypt.

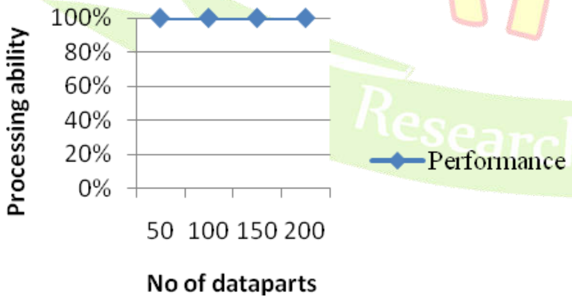

Figure 2: Processing ability of the proposed system

From Fig 2, the processing ability of the system remains the same even when no of data parts increase. Here a searching process is done and no of searched item is returned. This in turn means that even when the size of the data increases, the proposed system is able to process it with good efficiency.

Regarding an unauthorized access, the data owner should know that an unwanted access to big data is made. Hence, for each tenant's username and password a set of honey words is generated and stored in a database such that only one string is the correct password and the others are decoy passwords.

In Chaffing-with-a-password-model, the password will be split into character sets. For example, crypt123 is decomposed as 5-letters + 3-digit and replaced with the same composition crypt172. When an adversary tries to enter into the system with a honeyword, an alarm is triggered to notify the data owner about an unauthorized access. This in turn helps to alert the data owner and tenants to safeguard the data.

V. CONCLUSION

In this work we proposed a Multicloud environment to securely store big data in E-Commerce. This cloud storage provides robustness, availability and avoids single point failure. A part of a large data will not be fruitful for the adversaries. It also reduces encryption overhead as only certain data parts are encrypted .When an authorized tenant performs the analytical operation, the performance of the proposed system is good. If an unauthorized access is made, then the data owner is alerted by means of an alarm.

REFERENCE

1. Alizain, M.A., Pardede, E.; Soh, B.; Thom, J.A., "Cloud *Computing Security: From Single to Multi-clouds*," *HICSS*, pp.5490-5499, Jan. 2012.

2. Chang-Ji Wang; Xi-Lei Xu; Dong-Yuan Shi; Wen-Long Lin, "*An Efficient Cloud-Based Personal Health Records System Using Attribute-Based Encryption and Anonymous Multi-receiver Identity-Based Encryption*," *PGCIC*, pp.74-81, Nov. 2014

3. Cheng Hongbing; Rong Chunming; Hwang Kai; Wang Weihong; Li Yanyan, "*Secure big data storage and sharing scheme for cloud tenants*," in

China Communications, vol.12, no.6, pp.106-115, June 2015.

4. Chi Chen; Xiaojie Zhu;Peisong Shen; J.Hu;S.Guo;Z.Tari; Albert Y. Zomaya, *"An Efficient Privacy-Preserving Ranked Keyword Search Method"*, IEEE Transactions on Parallel and Distributed Systems, Jan 2015.

5. Erguler, I., "*Achieving Flatness: Selecting the Honeywords from Existing User Passwords*," IEEE Transactions on Dependable and Secure Computing, vol.PP, no.99, pp.1-1,2015

6. Ning Cao; Cong Wang; Li, Ming; Kui Ren; Wenjing Lou, "*Privacy-preserving multi-keyword ranked search over encrypted cloud data,*" 2011 Proceedings IEEE in INFOCOM, pp.829-837, April 2011

7. Kan Yang; Xiaohua Jia, "*Expressive, Efficient, and Revocable Data Access Control for Multi-Authority Cloud Storage*," IEEE Transactions on Parallel and Distributed Systems , vol.25, no.7, pp.1735-1744, July 2014

8. Ulusoy, H.; Kantarcioglu, M.; Thuraisingham, B.; Khan, L., "*Honeypot based unauthorized data access detection in MapReduce systems*," 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), pp.126-131, May 2015

9. Yakoubov, S.; Gadepally, V.; Schear, N.; Shen, E.; Yerukhimovich, A., "*A survey of cryptographic approaches to securing big-data analytics in the cloud,*" 2014 IEEE in High Performance Extreme Computing Conference (HPEC), vol., no., pp.1-6, Sept. 2014.

10. Zhi-Hong Tian; Bin-Xing Fang; Xiao-Chun Yun, "*An architecture for intrusion detection using honey pot,*" 2003 International Conference on in Machine Learning and Cybernetics, vol.4, no., pp.2096-2100 Vol.4, Nov. 2003