EFFECTIVE CONTENT SEARCH AND DE-DUPLICATION IN CLOUD DATA

[1]S. Lalitha , [2]N. Kamal Raj
[1]M.Phil Scholar, [2]Head of the Department (IT),
[1]Department of Computer Science
[1]Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India

Abstract: In today's environment every user wants to store and access their data at any time and at anywhere. Cloud storage system allows users to access their stored data using several virtual machines. Due to this scalable nature, the clouds are vulnerable and possible for several security and privacy issues. In order to provide the privacy and security for the encrypted data, a new framework is proposed. The proposed framework is named as "**PASS**" (**P**rivacy **and S**ecure data **S**earch and de-duplication"). The proposed **PASS** framework identifies security and privacy issues in cloud environment and eliminates the duplicate records at the client side. The **PASS** framework helps to protect the user search privacy and content security over encrypted data. Compared with the existing schemes, the scheme only need to check a small portion of ranked indexes in a results and, thus, greatly reduces the verification cost. The PASS scheme supports different multi-weighted keyword semantics over encrypted information and this also verifies the integrity of the order within the search result. The proposed system aims to achieve high security and privacy for cloud data with increased search efficiency, accuracy and time efficiency.

Keywords: Cloud Computing, De-Duplication, Record Matching, Document Clustering, Privacy and Security

1. INTRODUCTION

Cloud computing is an emerging technology which provides on demand services, which is based on virtualization, parallel, and distributed computing, utility computing, and service oriented architecture [1]. The most useful service which has emerged in the IT industry and the academic world is cloud computing. The uses of cloud computing include reduction in costs in capital expenditures, increased operational efficiencies, scalability, flexibility, immediate time to market, and many more [2]. The different cloud computing services are Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [3]. The service provided by cloud is subscription-based service. Anyone can get networked storage space and computer resources at any time. The biggest challenge in the adoption of cloud technology is data security and privacy. Cloud computing provides cloud storage as a service that stores and manages data for their clients. There is a risk when trusting the cloud provider to store important data (files) on the server. While utilizing different cloud services, need to consider some possibilities such that a user may want to change cloud provider, or the cloud provider may close the business, etc. This work focuses on privacy issues at the time of searching, and effective de-duplication is performed in the proposed system. Data protection includes securing data as it is transmitted to and stored in the cloud as well as granting the appropriate access rights regarding who can view the data. Privacy is the level of confidentiality provided to the user in a system. Privacy not only guarantees the fundamental confidentiality of company data but also guarantees the data's level of privacy. Privacy can be violated by the intentional release of private company information or through a misapplication of network rights. The threat of data privacy is high in the cloud than traditional technology due to the number of interactions between risks and challenges. These are because of the architectural or operational characteristics of the cloud environment.

The proposed system develops to design and analyze a new Fast and Reliable privacy and security Techniques for effective data search and deduplication on encrypted cloud data. The system also proposes a clustering method to perform the document similarity calculation and clustering for confidential data. The Confidentiality of the data search over the cloud by the end user is preserved and analyzed. Deployment & Integration of different algorithms and techniques simultaneously for providing Confidentiality, Privacy, and Authentication of data over secure Cloud. The proposed system aims to reduce the verification cost and search time over encrypted data. And this also aims to handle high volume of data, so the scalability will be considered. The system also has an objective to perform data dynamics and integrity in the private cloud. The major objectives such as providing high security for the outsourced cloud data, performing data clustering of encrypted data using tag concepts, allowing multi keyword based search over encrypted content, de-duplicating the outsourced data using indexing and tag verification and identifying duplication in software services also. Because, cloud duplicate data is one of the issues accompanying the rapid growth of data on the cloud and the growing need is to integrate data from heterogeneous sources for making the cloud storage less. So the effective document duplication detection from the encrypted cloud storage is the main aim of the system. The proposed system introduces a new secure framework for document duplicate detection and cluistering. The framework is named as PASS, which consist many algorithms and techniques, which are i) Improved Hidden vector encryption algorithm, this contains the set of processes such as (Setup, Encrypt, KeyGen, indexing , search and verification) ii). Adaptive Agglomerative Clustering (AAC), this used to group the documents based on its similarity nature. iii)

Dynamic Hash Tree with sporadic is used to update the indexing process. This will improve the search efficiency. iv) Bloom filter and bloom search for fast document search, this will adequately helps to handle huge number of clients in the common cloud environment.

2. PROBLEM DEFINITION:

One challenge in the cloud is finding that the relationship between documents will be normally concealed in the process of encryption, which will lead to significant search accuracy performance degradation. Also the volume of data in data centers has experienced a dramatic growth. This will make it even more challenging to design cipher-text search schemes that can provide efficient and reliable online information retrieval on large volume of encrypted data. The existing techniques [4][5][6][7] is failed to perform document security in terms of integrity and dynamics analysis, the indexing process was not completely studied for data search as well as data de-duplication [8][9]. The searching accuracy in fully encrypted dynamic data is low and the computational overheads are high. However, applying the privacy and security on the data and user information in the encrypted cloud data search system remains a very challenging task because of inherent security and privacy obstacles, including various strict requirements like the data privacy, the index privacy, the keyword privacy, and many others. From the survey [10], the issues and challenges is identified and gathered. , based on those observations, the proposed research developed a new framework for the cloud storage services.

3. PROPOSED SYSTEM

The proposed work identifies security and privacy issues for secure data management in cloud environment. An efficient privacy preserving data search and verification scheme is proposed to protect the user search privacy and content security over encrypted data. Compared with the existing schemes, the scheme only need to check a small portion of ranked indexes in a results and, thus, greatly reduces the verification cost. In the proposed system, explore supporting different multi-keyword semantics (e.g., weighted query) over encrypted information and checking the integrity of the order within the search result. In the proposed system, an adaptive agglomerative clustering (AAC) method is proposed to support more search semantics and also to meet the demand for fast cipher text search within a dynamic huge data environment. The proposed system aims to provide security and privacy for cloud data with increased search efficiency, accuracy and time efficiency.

3.1 Research contributions:

The goal of the research is to provide a solution to the research problems that have been stated in the literature. In cloud computing, data owners can share their files which are stored in cloud with a large number of authorized users, who may wish to retrieve only certain specific files they are interested in a given period. One of the most familiar methods to do is search word based search. The index creation time and the space used by the index is more in all existing methods. The proposed method uses clustering techniques for file retrieval which allows the users to search over encrypted data in a secure manner through keyword search and retrieval of the files with duplicate detection. The duplicate detection is based on relevance scores. This method enhances the system usability by enabling the deduplication. Since is does not send unwanted results, it also ensures the file retrieval accuracy. The experimental result shows the proposed method is efficient, secure and less time and space consuming technique. The traditional way to protect data is through data encryption before outsourcing to the cloud storage system. Even though the algorithms are public, the files encrypted under the encryption algorithms are secure because of keeping the symmetric key secret. As a result it is necessary to share the symmetric key secretly to the users in the cloud computing to avoid security issues. In all the existing systems, a single root node keeps the master key which is used to get other node keys. The most common problem with this method is that it is time consuming to derive the child nodes key and it is not secure. Considering the above problem, an efficient and secure key generation and crypto_search for cloud data sharing method is proposed. With the proposed IHVE algorithm, the encrypted data is searchable with the tag concepts. The set of algorithm and techniques used in proposed system such as Improved Hidden vector encryption: this contains the set of processes such as (Setup, Encrypt, KeyGen, indexing, searching and verification), Adaptive agglomerative clustering (AAC) is proposed to group the documents based on its similarity nature. Dynamic Hash Tree with sporadic is used to update the indexing process. This will improve the search efficiency. Bloom filter and bloom search for fast document search, this will adequately helps to handle huge number of clients in the common cloud environment. And Semantic calculation between terms is performed for data grouping and analysis.

The proposed data de-duplication and data search framework “PASS” has the several advantages such as It reduces the verification cost and time, this handles high volume of data and clusters the data based on the encrypted vectors. PASS framework performs multi-keyword search over encrypted data, this provides fast search. The overall process of the proposed system is depicted in the figure 1.0

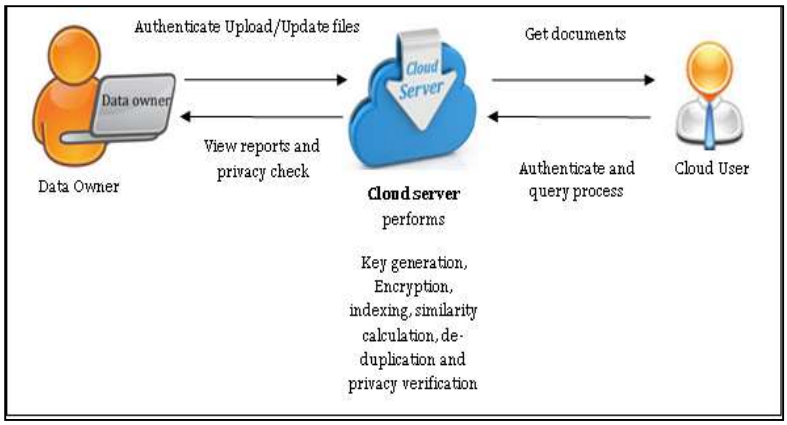

Figure **1.0 PASS** architecture

The users wish to retrieve only certain specific files they are interested, in a given period. To preserve data privacy, a basic solution is to encrypt data files and then upload the encrypted data into the cloud. This is normally done by cryptographic method and keeping the encryption key secret from the cloud service provider, in the proposed method this functionality is handled by the IHVE method. The generation and distribution of a symmetric key is important for maintaining confidentiality. The cloud users need to check whether the data are available in the cloud before any data usage such as computation or search over cloud data. So it is required to design an efficient file retrieval method that will overcome the above said problems. This can be performed using the PASS framework. The PASS is consisting with the cloud user and cloud service provider.

3.2 PASS framework

The proposed PASS framework consists of three types of users, which are the cloud server with Virtual machine (VM), cloud user and data owner. The detailed methodology of the proposed system is discussed below. The proposed PASS architecture consisted of different algorithms, this initially defines the overall algorithm of the proposed system, and then the individual process is explained,

Process steps:

1. Initiate cloud service provider CSP.
2. Register client C and generate a secret key Sk.
3. Get document D from the C and perform IHVE.
   a. Ed=IHVE(key, D,C), where Ed is the encrypted document
   b. Get hash index (HI) for every vector V.and store in the Hash index.
4. Upload Ed to the CSP.
5. Search over encrypted data:
   a. Get query Q from Client.
   b. Search the Q in the vector index from the Bloom HI.
   c. Calculate similarity S=Sim(Q,Edi)
   d. Retrieve the files based on the similarity S.
   e. Return Edi.
6. Deduplication process:
   a. Sort files based on the similarity score S.
   b. Match Vector list V for every Edi in CSP using the step 5.c
   c. Duplicate threshold T. if (SEdi>T)
   d. Return Edi as duplicate and add to document hash with index I.
7. Decrypt and end.

The Owner initially creates the file which is needed to be uploaded in the Cloud Server then creates a document score table which consist the vectors as a tag and the score for each file for that corresponding search word. Then the file is encrypted and encrypted file and the score table is uploaded to the server. Only the Owner knows the files that are being uploaded into the Cloud server and he has the overall control over the system. The proposed PASS framework consist of the IHVE based encryption and searching process.

Phase 1: Cloud infrastructure and authentication

Network infrastructure creation with n number of servers and clients is the first step. The module creates the following interfaces such as Storage server with numerous CSP (Cloud service Providers) and Client. The authentication phase defines the security and authority to access the above user types, for example every client should be authenticated before accessing the resources in the storage cloud. Only the authenticated persons can upload and download the files. For this process user should register with all basic information. The file should be secured before transmitted in to the storage servers. The first module consists of the following sub processes.

CSP: the storage server has the responsible to respond for the client request. The allocations of server configurations are performed in this module.

Client: the clients are separated into two types, the data owner and cloud user.

Phase 2: Data upload process:

This module will be performed by the data owners after the successful authentication. The data owner can upload a text document with the accessibility specification. The access specification is nothing but specifying the document as private or public.

Phase 3: Key generation and encryption Process:

The third module generates the secret key for every document and encrypts using IHVE. This performs the following process.

Keygen: The public and secret key pair are created by this phase

From the Master secret key MSK→ decryption key SK will be generated.

Encryption: the next process is document encryption process, which takes message M as input and public key.

This module performs the above and stores in the cloud storage.

Phase 4: Indexing and verification:

The IHVE scheme performs the index process and verification process at the time of encryption and decryption. This process simplifies the data retrieval delay and difficulty in encrypted text over the encrypted cloud storage.

Phase 5: query processing

The query processing step involved with user data request process. This detects the weighted terms and finds the rule to access the generated result documents. While retrieving the document, the system finds the relevant score calculation step to retrieve appropriate documents from the cloud storage.

Phase 6: Tag and index Creation for data de-duplication:

For the fast retrieval and effective duplication detection over encrypted content, the system proposes a new dynamic index tree and tag creation method for frequent data search. This process stores the previous search result in the index and performs the new search after the verification in hash table. This type of process eliminates the redundant search and time delay also.

This module uses the vector method to verify the similar data without decryption. The system uses the IHVE algorithm and AAA for effective data search and de-duplication algorithm for redundant content avoidance.

Phase: 7 content decryption

After the successful search the encrypted contents are identified by the user query. The retrieved documents will be decrypted according to the users master secret key and decrypted key. The documents are decrypted after the verification of user keys and owners privacy.

Phase 8: Performance analysis and comparison

The final module generates different type of charts and graphs to show the proposed system performance. This includes the encryption delay, key generation, indexing and verification delays along with the privacy risk measures and these details will be compared with the existing techniques.

4. IMPLEMENTATION AND RESULTS

The proposed model uses the Distributed file System to store and retrieve the top-n files using the index which are distributed in different cluster machine to perform the parallel operation. The experimental setup was implemented in windows 10 along with cloud infrastructure with 2 clusters, cluster1 and cluster2. Here the basic idea is to transfer the file between the two clusters. The implementation was done using C#.net. The experiment was carried out between the two cluster, the cluster 1 acts as the client medium and cluster2 running in 1st core of the Intel dual core processor acts as the server.

The proposed system works efficiently and produces optimum solution even if the cloud contains very large number of files. The index creation time and storage is decreased. The performance is compared with several parameters to prove the effectiveness of the proposed system based on the no of keywords in the file.ie based on file size in terms of no of keywords.

4.1 Time Complexity

The time complexity of the proposed algorithm is O (n), because the performance will grow linearly and in direct proportion to the total number of input file collection.

Table 1.0 time complexity analysis table

File size	*Encryption time*	Vector and tag process time	Search time
50	*43*	60	8
100	*78*	98	15
300	*120*	148	23

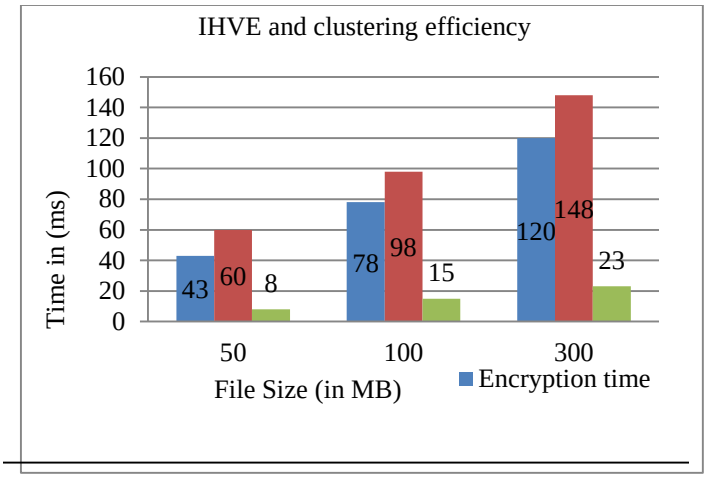

Figure 2.0 time analysis for different process in the proposed system

The time taken for different process is given in the table 1.0 and the graph shows the time consumption for encryption, vector and tag selection and search time. This shows, the search time is reduced due to the effective vector and tag generation process.

4.2 Search efficiency:

The proposed system measures the search efficiency by performing different tasks with different set of keywords and files.

Table 2.0 search efficiency table

Experiment	**Total file count**	**Successfully retrieved files using the proposed system**
1	20	20
2	30	28
3	50	48

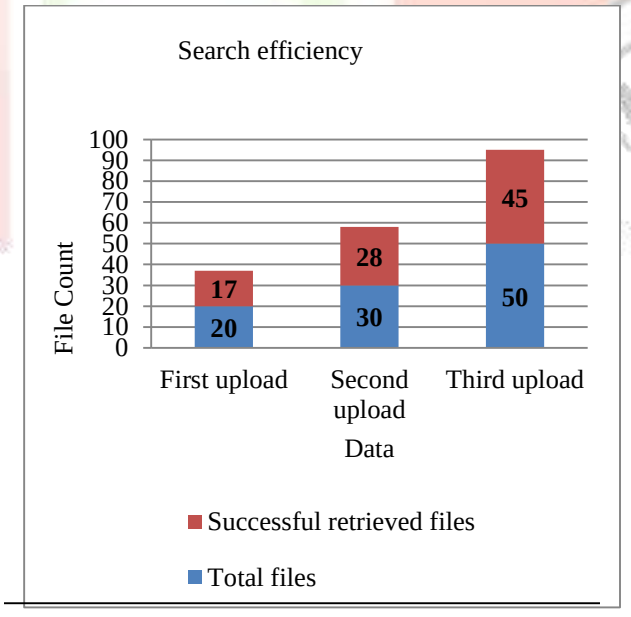

Figure 3.0 search efficiency of the proposed system

The figure 3.0 shows the search efficiency of the proposed system, which experimented for three types of process. The first upload processes begin with the total 20 files and 17 are extracted correctly. Likewise, for every upload the accurate result count is analyzed.

Table 3.0: Comparison of Cryptographic Process accuracy

Accuracy (%)	***Server side duplication***	***HVE***	***IHVE***
	73.151	90.687	96.852

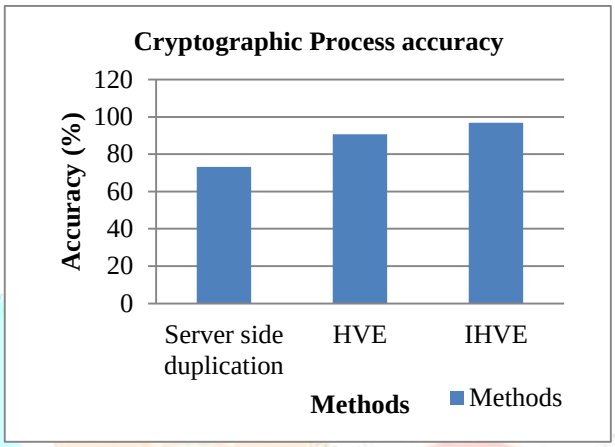

Figure 4.0 Cryptographic Process efficiency in terms of retrieval and search accuracy

Experiments have shown in table and figure 4.0 that the proposed IHVE algorithm is effective for the search over encrypted document in less time with reduced error in terms of search accuracy. The proposed method can also be used in the development of tools for the analysis of the dynamics of scientific and technical expertise in the collections of electronic files.

CONCLUSION:

The focus of this paper is to provide the data mining services to the detection of duplicate documents in the cloud encrypted source. This allows the user to search the content from the cloud over encrypted content based on the user requesting the information without decrypting it. Several methods have been proposed to perform cloud data security and search over encrypted content with duplication, but still needs some enhancement for the encrypted data deduplication. It is hypothesized that because data mining approach can measure similarity between two outsourced files and effectively duplicates by membership values and because cloud encrypted documents contain complex and dynamic information for clustering, Adaptive agglomerative clustering will be more effective with efficient vector based tag selection, duplicate detection and document clustering. The major objective of this paper is to build up an enhanced duplicate file detection technique along with the similarity detection by clustering documents with high clustering accuracy, enhanced vector tag selection and also to decrease the convergence time and the number of iterations by duplicate detection. The evaluation results shows that the overall performance of the proposed approach provides encouraging results after applying PASS with the IHVE and AAA in terms of accuracy values.

References:

[1]. Xu, Dong. "Cloud computing: An emerging technology." In *Computer Design and Applications (ICCDA), 2010 International Conference on*, vol. 1, pp. V1-100. IEEE, 2010.

[2]. Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges." *Journal of internet services and applications* 1, no. 1 (2010): 7-18.

[3]. Bhardwaj, Sushil, Leena Jain, and Sandeep Jain. "Cloud computing: A study of infrastructure as a service (IAAS)." *International Journal of engineering and information Technology* 2, no. 1 (2010): 60-63.

[4]. Kandukuri, Balachandra Reddy, and Atanu Rakshit. "Cloud security issues." In *Services Computing, 2009. SCC'09. IEEE International Conference on*, pp. 517-520. IEEE, 2009.

[5]. Feng, Deng-Guo, Min Zhang, Yan Zhang, and Zhen Xu. "Study on cloud computing security." *Journal of software* 22, no. 1 (2011): 71-83.

[6]. Ng, Wee Keong, Yonggang Wen, and Huafei Zhu. "Private data deduplication protocols in cloud storage." In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 441-446. ACM, 2012.

[7]. Stanek, Jan, Alessandro Sorniotti, Elli Androulaki, and Lukas Kencl. "A secure data deduplication scheme for cloud storage." In *International Conference on Financial Cryptography and Data Security*, pp. 99-118. Springer, Berlin, Heidelberg, 2014.

[8]. Xu, Jia, Ee-Chien Chang, and Jianying Zhou. "Weak leakage-resilient client-side deduplication of encrypted data in cloud storage." In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, pp. 195-206. ACM, 2013.

[9]. Leesakul, Waraporn, Paul Townend, and Jie Xu. "Dynamic data deduplication in cloud storage." In *Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on*, pp. 320-325. IEEE, 2014. S.

[10]. Lalitha and N. Kamal Raj " A Survey on Data De-Duplication Methods in Cloud Storage System" in International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) Vol. 5, Issue 7, July 2017: 13431- 13438.