Duplicate Detection in Cloud Data Storage using various Encryption Algorithms and Techniques

S. Senthil Kumar,

Ph.D. Research Scholar,
Department of Computer Science,
Kongunadu Arts and Science College, (Autonomous), Coimbatore,
Tamil Nadu,
India.
E-mail: ssksnsmca@gmail.com

Abstract:

Cloud computing is a new emerging technology which provides on demand services that is based on virtualization, parallel, and distributed computing, utility computing, and service oriented architecture. For many years in the past, 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. The different cloud computing services are Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The service provided by cloud is subscription-based service. Anyone can get networked storage space and computer resources at any time.

Keywords: Cloud Computing, Encryption, Duplication

INTRODUCTION

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 management 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.

To overcome the issues with several new features are used.

- The research aimed to design an effective data management scheme with privacy and security considerations, to provide cloud user the confidence to store and retrieve sensitive data on cloud storage.
- To improve the data retrieval efficiency in the encrypted data.
- To improve privacy and response time with less compromise on the client site communication and computation cost and time.

Cloud computing encompasses many technologies and tools, due to its virtual nature, there are several security and performance issues arises. Some security issues and drawbacks related to the research are given below.

- Unauthorized Access to servers and applications
- Data duplication occurs in the cloud. Sometimes this is happen due to the attacker.
- Virtual Machine Security and authentication
- Data Privacy, Data Integrity and Data Security need more computation
- Security policy and fulfillment

Related Work

Duplicates can occur in numerous situations, for instance when a large database is updated by an external source and registry numbers are not accessible or are in error. Organizations are often confronted with the need to identify duplicate records present in huge databases. In a population register, there is a chance of some individual entities being listed under two or more registry numbers. Some information such as, name, address and date-of-birth may be necessary to identify the duplicates. Address or date-of-birth information is needed additionally as names do not uniquely identify. The identification of duplicates is difficult when names, addresses, and dates-of birth contain typographical errors. Ahead of mining the accurate models, the data from the relevant sources must be collected, integrated, cleaned and pre-processed in different ways. The merging of data from multiple databases into a single relation can often result in several duplicate records. These records are not syntactically identical, but, the same real-world entity is represented by them. To produce data of sufficient quality for mining, it is important to appropriately merge these records and the information they represent. Names such as record linkage, de-duplication, merge/purge, object identification, identity uncertainty, hardening soft information sources and more are also employed to denote this problem. The task of quickly and accurately identifying the records corresponding to the same entity from one or more data sources is known as record linkage [1]. The term approximate or fuzzy duplicates refer to the tuples which are somehow different but describe the same real world entity. The elimination of fuzzy duplicates in any database is necessary and yet it is vital in the data integration and analytical processing domains that require accurate reports/statistics [2]. The authors have utilized an approach for de-duplication that makes use of a fuzzy logic framework. The fuzzy inference system was optimized with the aid of the Bayesian Optimization Algorithm, a class of Estimation of Distribution Algorithms that are capable of learning complex multivariate relations of bounded order. Breeder genetic algorithms, utilized in the science of livestock breeding, were the motive behind the proposed class of algorithms. In Digital Libraries due to diverse sources of books that are spread across various parts of the country, duplicates could arise between scanning points. The Duplication of the books can be identified using only metadata of a book. If the metadata is missing, incorrect, abbreviated or incomplete it makes the duplicate detection all the more difficult.

There are several problems threatens the cloud environment by above issues. To protect privacy of data and oppose unsolicited accesses in the cloud and beyond it, sensitive data, for instance, e-mails, personal health records, photo albums, tax files, and so on, may have to be encrypted by data owners before outsourcing to the commercial public cloud; this, however, the main obstacle in that is it follows the traditional search service which can only search based on plaintext keyword [3]. The insignificant solution of downloading all the data and decrypting locally is clearly impractical, due to the large amount of bandwidth cost in cloud scale systems. Every data in the cloud will be secure and private, So exploring privacy preserving and secure search service over encrypted cloud data is of great importance. And the duplication detection at the client side is also important. The content duplication detection and searching process on encrypted file is performed by several authors in the literature, however the problem is these processes are particularly challenging when it is associated with the performance oriented issues.

Cloud data indexing is provided for fast search, but the priorities of all the data files is kept same so that the cloud service provider and third party remains unaware of the important files, thus, maintaining privacy of data. Ranked search can also elegantly eliminate unnecessary network traffic by sending back only the most relevant data, which is highly desirable in the “pay-as-you-use” kind of cloud pattern. For privacy protection, such ranking

operations however, should not leak any keyword related information. Besides, to improve search result accuracy as well as to enhance the user searching experience, it is also necessary for such ranking system to support multiple keyword searches, as single keyword search often produces many results than the expectation. Along with the privacy of data and efficient searching schemes, real privacy is obtained only if the user's identity remains hidden from the Cloud Service Provider (CSP) as well as the third party user on the cloud server.

METHODOLOGY

Data Owner

The Owner initially creates the file which is needed to be uploaded in the Cloud Server then creates a document score table which consists the vectors as a tag and the score for each file for that corresponding search word [1]. 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. Fig (A) shows the flow diagram of owner process.

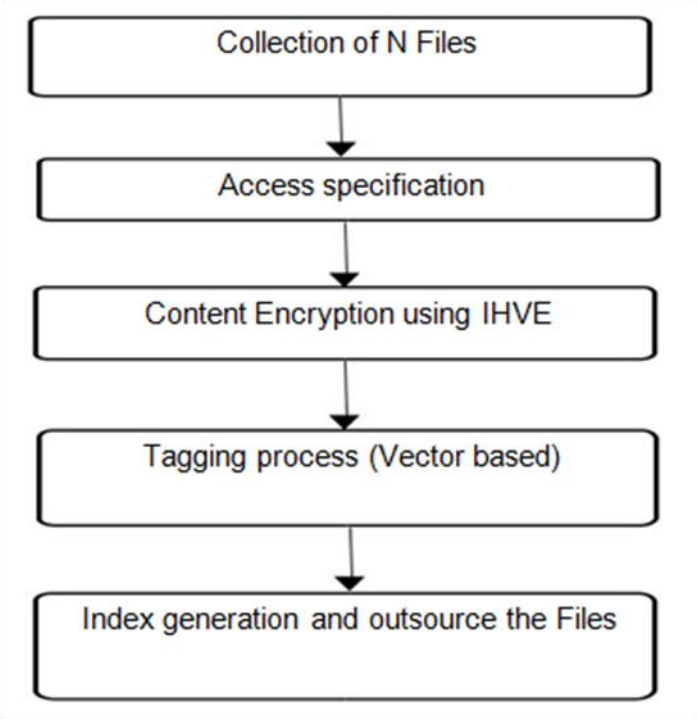

Fig (A): Owner process

The data owner initially collects the files from the local system, which needs to be outsourced to the cloud. In the proposed system, the file can be any format, like document .txt or a software. After data collection the data owner should specify the file access type such as Private or public. If the data type is specified as Private, then the data will be retrieved only by the authorized user.[2] This private data will be secured by the encryption process.

Cloud Users

The main role of the user is to search for the files needed using the search words present in the file. The search bar is created using the Query deployment module in which the User will enter the query.[3] The result i.e., the top n files will be sent back to the user for the query entered by them. Fig (B) shows the flow diagram of owner process.

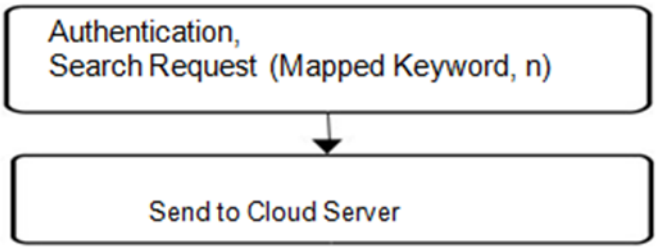

Fig (B) User process

Cloud Server

The server takes care of the operations such as the storing and extraction of the files which will be discussed below. The server acts as the host to which all the users can get connected. The server as shown in the center handles all the information of input, processing and responding [4].The client system doesn't need any kind of extra system to handle all information. The user side can simply receive the information needed. All the data stored in the cloud storage are in the encrypted form by encrypting using the symmetric encryption algorithm. This enhances the security for the data stored in it. Fig (C) shows the flow diagram of owner process.

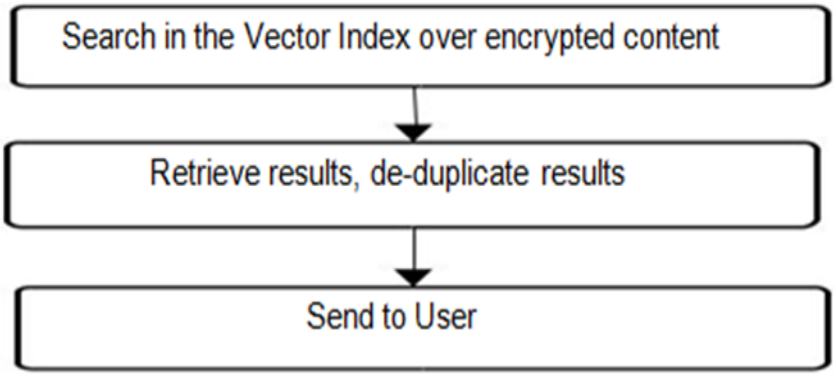

Fig (C): Cloud Server process

A secure indexed file retrieval method uses four main steps. Key generation, tag index generation algorithm, Encryption algorithm and search algorithm. These algorithms are implemented in two phases' setup and retrieval.

Document similarity and clustering for duplicate file detection:

The tag vectors for each file are generated by reading a line at a time and splitting the data file into vectors and remove any vectors present and stem the resulting tokens. For each vector in the file the vector count is also generated.

Cloud Services:

1· Cloud Software as Service Here, any user can make use of the cloud provider's running applications at any given point of time.

2· Cloud Platform as Service It is the ability to deploy the applications developed by the user using any programming languages or tools made available by the cloud service provider on the cloud infrastructure.

3· Cloud Infrastructure as Service In Infrastructure as a service, the user can deploy and execute the software. The user will also have the provision for processing, storing and other basic computing resources.

Cloud Deployment Models:

1. Public cloud: Any user with an internet connection can use the cloud space in Public Cloud.

2. Private Cloud: A Private Cloud is exclusive for a specific group or organization and the usage of this cloud is restricted to a group or organization alone.

3. Community Cloud: Two or more organizations which have common cloud requirements share a Community Cloud.

4. Hybrid Cloud: A hybrid cloud is basically a blend of at least two clouds, which could be a combination of public, private, or community.

IMPORTANCE OF RECORD DUPLICATE DETECTION IN CLOUD

Cloud data storage plays a significant role in the current scenario. A number of organizations require quality data. Data quality problems arise with the constantly increasing quantity of data stored in real-world databases that are assured by the vital data cleaning process. Data quality problems are encountered in the single data collections, like the files and databases. For example, owing to misspellings during data entry, unintentionally omitted information or other invalid data or due to the integration of multiple data sources in data warehouses results in significant increase in the need for data cleaning[5].

Data cleaning deals with the detection and removal of errors and inconsistencies from the data to improve the quality of data. Data cleaning plays a significant role in the process of data mining. It is necessary to enrich the quality of data in a data warehouse prior to the data mining process. Numerous data cleaning techniques are being employed for diverse purposes.

The fundamental element of data cleaning is usually termed as duplicate record identification, but when it comes to the cloud, the data can't be easily deleted or de-duplicated. The duplicate record detection is the process of identifying the record pairs signifying the same entry[6]. The process of duplicate detection is preceded by a data preparation stage which includes several steps such as data parsing, indexing, tagging and a data transformation, during which data entries are stored in a uniform manner in the database resolving the structural heterogeneity problem. Data preparation is also described using the term ETL (Extraction, Transformation, Loading)

Duplicate Record Detection

Multiple versions of the same record are often accumulated when databases are constructed from multiple sources. The task of detecting these different versions is known as record de-duplication. All records that contain exactly or approximately the same data in one or more fields are identified in the process of duplicate detection[7]. The problem of identifying syntactically different records that describe unique entities is denoted by all terms such as record linkage, duplicate detection and more. There are several approaches for solving duplicate detection problem. Some of the approaches are Probabilistic Matching Models, this model uses a Bayesian approach to classify record pairs into two classes.

Supervised Learning

The supervised learning depends on the presence of training data in the form of record pairs, prelabeled as same or not. Cochinwala et al used a popular CART algorithm generates classification and regression trees. A linear discriminant algorithm is also used to generate linear combination of the parameters for separating the data according to their classes and a .vector quantization method which is a generalization of nearest neighbor algorithms.

Unsupervised Learning

The idea behind this learning is to avoid manual labeling of the comparison vectors through clustering methods.

Active Learning based Techniques

Unlike an ordinary learner that trained using a static training set , this technique uses active learner which actively selects subsets of instances from unlabeled data which, when labeled deliver the highest information gain to the learner.

Distance Based learning

Supervised learning and active learning techniques are not appropriate in the absence of training data. One way of avoiding the essential of training data is to define a distance metric for records which does not need the training data. In this approach the similar records is matched by using the distance metric and an appropriate matching threshold.

A number of researchers belonging to different groups, including databases, and machine learning have been studying this problem. The duplicate records under a single representative record are detected and then grouped or clustered in the duplicate record detection.

De-duplication can be performed for a group of databases or for a single database that contains duplicate records. An 'error free' approach in the data warehouse is known as data quality. It is essential to enrich the quality of data through data cleaning methods[8].

Numerous data cleaning techniques are being employed for diverse purposes. Similarities among records and fields are identified using 'Similarity. 'Duplicate elimination functions' are employed to identify if two or more records signify the same real world objects. Data cleaning methodologies which are in existence have been employed to recognize the missing values, record and field similarities and duplicate elimination[7]. As it is not possible to assume a unifying set of standards for various data sources, these issues are unavoidable. With increase in the size of the database, the

problem intensifies[9]. This is due to the huge amount of computational resource required for the examination and removal of duplicate records .

Duplicates can occur in numerous situations, for instance when a large database is updated by an external source and registry numbers are not accessible or are in error. Organizations are often confronted with the need to identify duplicate records present in huge databases. In a population register, there is a chance of some individual entities being listed under two or more registry numbers. Some information such as, name, address and date-of-birth may be necessary to identify the duplicates. Address or date-of-birth information is needed additionally as names do not uniquely identify. The identification of duplicates is difficult when names, addresses, and dates-of- birth contain typographical errors[10].

Ahead of mining the accurate models, the data from the relevant sources must be collected, integrated, cleaned and pre-processed in different ways. The merging of data from multiple databases into a single relation can often result in several duplicate records[11]. These records are not syntactically identical, but, the same real-world entity is represented by them. To produce data of sufficient quality for mining, it is important to appropriately merge these records and the information they represent. Names such as record linkage, de-duplication, merge/purge, object identification, identity uncertainty, hardening soft information sources and more are also employed to denote this problem. The task of record detection, which are similar or nearly similar among other entities should be detected quickly and accurately. This process is commonly known as record linkage. The term approximate or fuzzy duplicates refer to the tuples which are somehow different but describe the same real world entity. The elimination of fuzzy duplicates in any database is necessary and yet it is vital in the data integration and analytical processing domains that require accurate reports/statistics[12].

In the IT security space cloud architectures are growing rapidly[13]. The common cloud providers present in the market are Amazon, Microsoft, Google, IBM, Oracle, Eucalyptus, VMware, Citrix, Sales force and Rack space. The cloud providers offer different types of services

❖ Amazon: Amazon Web Services including the Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3), etc. Varied applications are developed using highly scalable, flexible, available computing platform provided by Amazon.

❖ Google: Google App Engine It provides application programming interfaces for the storing of data, Google accounts, and manipulation of image and e-mail services.

❖ Microsoft: Windows Azure Platform. This platform is a collection of cloud technologies. It offers a definite set of services to specific application developers.

Encryption Algorithms and Techniques for Duplicate Detection in Cloud Data Storage:

Many encryption algorithms and techniques are:

i) Improved Hidden vector encryption algorithm, this contains the set of processes such as (Setup, Encrypt, KeyGen, indexing , search and verification)
ii) Hierarchical Adaptive Agglomerative Clustering (HAAC), this used to group the documents based on its similarity nature.
iii) Dynamic Scheduling Algorithm 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.

Improved Hidden vector encryption algorithm

Improved Hidden vector encryption (HVE) scheme consists of the following four probabilistic polynomial-time algorithms:

• Setup(1k , Σ, L): on input a security parameter 1k , an alphabet Σ, a vector-length L, the algorithm outputs a public key PK and master secret key MSK.

• Encryption(PK, −→v , M): on input a public key PK, a message M, a vector v ∈ Σ∗ L where Σ∗ denotes Σ ∪ {∗}, the algorithm outputs a ciphertext CT .

• KeyGen(MSK, −→x): on input a master secret key MSK , a vector −→x ∈ ΣL, the algorithm outputs a decryption key SK.

• Decryption(CT, SK): on input a ciphertext CT and a secret key SK, the algorithm outputs either a message M or a special symbol ⊥.

Hierarchical Adaptive Agglomerative Clustering :

Algorithm Hierarchical Adaptive Clustering is Input:

- the set X = {O1,...,On} of m-dimensional previously clustered objects;

- the set X = {O 1,...,O n} of (m+s)-dimensional extended objects to be clustered; O i has the same first m components as Oi;

- the metric dE between objects in a multi-dimensional space; - the number p of desired clusters; - K = {K1,...,Kp} the previous partition of objects in X.

Output:

- the new partition K = {K 1,...,K p} for the objects in X .

Begin

For all clusters Kj ∈ K do

Calculate Corej ← (StrongCorej ≠ ∅)?StrongCorej : W eakCorej

Calculate OCorej ← Kj \ Corej

EndFor

C ← ∅ // the current cluster set

For i = 1 to p do

If Corei ≠ ∅ then C ← C ∪ {Corei}

EndIf

For all O ∈ OCorei do C ← C ∪ {O} //add a singleton to C

EndFor

EndFor

While | C |> p do

(Cu∗ , Cv∗) ← argmin(Cu,Cv)dE(Cu, Cv)

$C_{new} \leftarrow Cu* \cup Cv* \; C \leftarrow C \setminus \{Cu* , Cv* \} \cup \{C_{new}\}$

EndWhile

$K \leftarrow C$

End.

Dynamic Scheduling Algorithm with sporadic

begin

· let t be the current clock time

· let the occurring task be a 2-uplet (C,d) where d=deadline,

C=processing time

· let L_p be the current list of periodic tasks

· let L_s be the current list of accepted sporadic tasks

· let I be the list of semaphores accessed at t

· construct the EDL_m schedule from L_p and compute the list of idle time periods,

· find the entry k in L_s such that $d < d_k$

· temporarily consider (C,d) as inserted prior to the entry k in L_s and be the current entry

.Q_0:=0

· for each entry j in L_s from the current entry to the last entry

do

· calculate O(t.dj) from 1

· Qj==Qj_1+Cj

· if (Q_j > $O(t._{dj})$) then

return (non accepted) end if

end do

· insert (C,d) in L_s

· return (accepted)

end

Bloom filter and bloom search Algorithm:

Bloom filter Algorithm:

Choose a ballpark value for n

Choose a value for m

Calculate the optimal value of k

Calculate the error rate for our chosen values of n, m, and k. If it's unacceptable, return and change m; otherwise we're done.

Bloom Filter Algorithm Working:

The bloom has to do with, as we have seen, two things. One is when an element is added and another when we check if an element is present in a set.

Adding elements to the set

When we want to add Elements to the set, they are hashed and the bits at the positions of the hashes are set to 1.

Suppose we want to add the string filter to the set. We get some hash values, say as follows.

p(filter) % 10 = 1

p(filter) % 10 = 4

p(filter) % 10 = 7

The positions 1, 4 and 7 in bloom filter now have their bits set to 1 and the rest are zero.

CONCLUSION

This paper reviews how the user search the content from the cloud over encrypted content based on the user requesting the information without decrypting it. Several encryption algorithms and 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 de-duplication. The major objective of this paper is to build up an enhanced duplicate file detection technique along with the similarity detection by clustering documents.

REFERENCES

[1] Danny Harnik, Benny Pinkas, Alexandra Shulman-Peleg: Side Channels in Cloud Services: Deduplication in Cloud Storage. IEEE Security & Privacy 8(6): 40-47, 2010

[2] Dave Russell: Data Deduplication Will Be Even Bigger in 2010, Gartner, 8 February 2010

[3] Min, J., Yoon, D., & Won, Y. (2011). Efficient deduplication techniques for modern backup operation. IEEE Transactions on Computers, 60(6), 824-840.

[4] Stanek, J., Sorniotti, A., Androulaki, E., & Kencl, L. (2014, March). A secure data deduplication scheme for cloud storage. In International Conference on Financial Cryptography and Data Security (pp. 99- 118). Springer, Berlin, Heidelberg.

[5] Ng, W. K., Wen, Y., & Zhu, H. (2012, March). Private data deduplication protocols in cloud storage. In Proceedings of the 27th Annual ACM Symposium on Applied Computing (pp. 441-446). ACM.

[6]Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: a survey. Knowledge Data Eng IEEE Trans. 2007;19(1):1–16. DOI: http://dx.doi.org/10.1109/TKDE.2007.250581.

[2]Qi X, Yang M, Ren W, Jia J, Wang J, Han G, Fan D. Find duplicates among the PubMed, Embase, and Cochrane Library databases in systematic review. PLOS One.2013. 8(8):e71838. DOI: http://dx.doi.org/10.1371/journal.pone.0071838

[7]. 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.

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

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

[10]. 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.

[11]. 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.

[12] 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.

[13] 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.