Data Compression Technique for Eliminating Duplicate Copies of Repeating Data in Cloud Storage

Sangeetha. B[1], Mr. E. S. K. Vijay Anand.[2]
[1, 2] G.K.M. College of Engineering and Technology, Chennai, Tamilnadu, India

Abstract- *Cloud computing means retrieve and storing information and programs over the Internet instead of your computer's hard drive. To protect confidentiality of the perceptive data while supporting de-duplication data is encrypted by the projected convergent encryption method before out sourcing. It makes the first attempt to properly address the problem of authorized data deduplication. We also present some new deduplication constructions supporting authorized duplicate in cloud using symmetric algorithm. Data deduplication is one of the techniques which used to solve the repetition of data. The deduplication techniques are commonly used in the cloud server for reducing the space of the server. To prevent the unauthorized use of data accessing and generate duplicate data on cloud the encryption technique to encrypt the data before stored on cloud server.*

Keywords- re-duplicate, empower duplicate validity, combination cloud

I. INTRODUCTION

Cloud computing provides seemingly unlimited "virtualized" resources to users as services across the whole Internet, while hiding platform and implementation details. In cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever-increasing volume of data. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

1. FEASIBILITY STUDY

A Feasibility Study is the analysis of a problem to determine if it can be solved effectively. The results determine whether the solution should be implemented. This activity takes place during the project initiation phase and is made before significant expenses are engaged.

Definition of Feasibility Stud

A feasibility study is an evaluation of a proposal designed to determine the difficulty in carrying out a designated task. Generally, a feasibility study precedes technical development and project implementation. A feasibility study looks at the viability of an idea with an emphasis on identifying potential problems and attempts to answer one main question: Will the idea work and should you proceed with it?

Five common factors (TELOS)

- Technology and system feasibility
- Economic feasibility
- Legal feasibility
- Operational feasibility
- Schedule feasibility

Technology and system feasibility

The assessment is based on an outline design of system requirements in terms of Input, Processes, Output, Fields **,**Programs, and Procedures. This can be quantified in terms of volumes of data **,**trends, frequency of updating, etc. in order to estimate whether the new system will perform adequately or not this means that feasibility is the study of thebased in outline**.**

Economicfeasibility

Economic analysis is the most frequently used method for evaluating the effectiveness of a new system. More commonly known as cost/benefit analysis the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs. If benefits

outweigh costs, then the decision is made to design and implement the system. An entrepreneur must accurately weigh the cost versus benefits before taking an action. Time Based.

Legal feasibility

Determines whether the proposed system conflicts with legal requirements, e.g. a data processing system must comply with the local Data Protection Acts.

Operational feasibility

Is a measure of how well a proposed system solves the problems, and takes advantages of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development**.**

Schedule feasibility

A project will fail if it takes too long to be completed before it is useful **.**Typically this means estimating how long the system will take to develop, and if it can be completed in a given time period using some methods like payback period. Schedule feasibility is a measure of how reasonable the project timetable is. Given our technical expertise, are the project deadlines reasonable? Some projects are initiated with specific deadlines. You need to determine whether the deadlines are mandatory or desirable**.**

II. PROPOSED SYSTEM

1. ARCHITECTURE

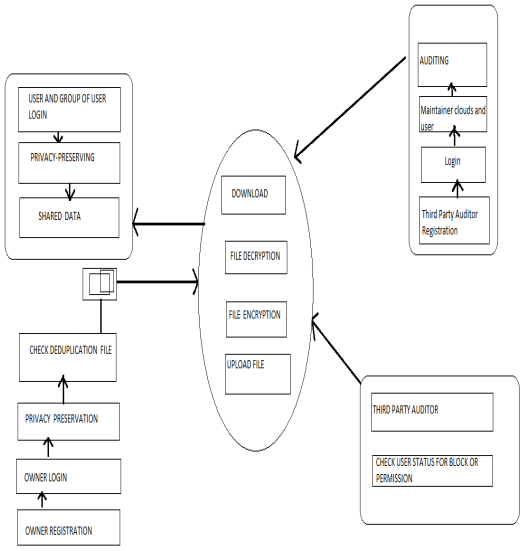

Fig 1. System Architecture

In using advanced deduplication system supporting authorized duplicate check. In this new deduplication system, a hybrid cloud architecture is introduced to solve the x problem. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server instead. In this way, the users cannot share these private keys of privileges in this proposed construction, which means that it can prevent the privilege key sharing among users in the above straightforward construction. To get a file token, the user needs to send a request to the private cloud server. The private cloud server will also check the user's identity before issuing the corresponding fill token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading this file. Based on the results of duplicate check, the user either uploads this file or runs POW.

Proposed System Advantage

In the proposed System Benefit's,

- Reducing the Storage Space
- Faster Recoveries
- Effectively increase network bandwidth
- Delete the duplicate files.
- High Security

2. MODULE DESCRIPTION

- ❖ Authorization control creation and Key Generation
- ❖ Owner uploading and Built Hybrid Cloud
- ❖ Detect Deduplication
- ❖ Key Exchanging
- ❖ Verification and File Retrieving

2.1 Authorization control creation and Key Generation

Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

Unauthorized users without appropriate privileges or file should be prevented from getting or generating the file tokens for duplicate check of any file stored at the S-CSP. In system, the S-CSP is honest but curious and will honestly perform the duplicate check upon receiving the duplicate request from users. The duplicate check token of users should be issued from the private cloud server in our scheme.

It requires that any user without querying the private cloud server for some file token, he cannot get any useful information from the token, which includes the file information or the privilege information.

2.2 Owner uploading and Built Hybrid Cloud

In this new deduplication system, a hybrid cloud architecture is introduced to solve the problem. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server instead. In this way, the users cannot share these private keys of privileges in this proposed construction, which means that it can prevent the privilege key sharing among users in the above straightforward construction. To get a file token, the user needs to send a request to the private cloud server.. To perform the duplicate check for some file, the user needs to get the file token from the private cloud server. The private cloud server will also check the user's identity before issuing the corresponding file token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading this file. Based on the results of duplicate check, the user either uploads this file or runs POW.

2.3 Detect Deduplication

Convergent encryption provides data confidentiality in deduplication. A user derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a *tag* for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness property holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side.

2.4 Key Exchanging

The private keys for the privileges are managed by the private cloud, the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively. The private cloud server will also check the user's identity before issuing the corresponding file token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading this file.

2.5 Verification and File Retrieving

A symmetric key x for each user will be select and set of keys will be sent to the private cloud. An identification protocol equals to proof and verify is also defined, where Proof and Verify are the proof and verification algorithm respectively. In each user U is assumed to have a secret key *to* perform the identification with servers.

Assume that user U has the privilege set PU. It also initializes a POW protocol POW for the file ownership proof. The private cloud server will maintain a table which stores each user's public information PKU and its corresponding privilege.

It first sends a request and the file name to the S-CSP. Upon receiving the request and file name, the S-CSPwill check whether the user is eligible to download file. If failed, the S-CSP sends back an abort signal to the user to indicate the download failure. Otherwise, the S-CSP returns the corresponding cipher text CF. upon receiving the encrypted data from the S-CSP, the user uses the key kef stored locally to recover the original file.

3. JAVA ARCHITECTURE

Java architecture provides a portable, robust, high performing environment for development. Java provides portability by compiling the byte codes for the Java Virtual Machine, which is then interpreted on each platform by the run-time environment. Java is a dynamic system, able to load code when needed from a machine in the same room or across the planet.

3.1 JVM

The Java Virtual Machine is the cornerstone of the Java platform. It is the component of the technology responsible for its hardware- and operating system independence, the small size of its compiled code, and its ability to protect users from malicious programs. The Java Virtual Machine is an abstract computing machine. Like a real computing machine, it has an instruction set and manipulates various memory areas at run time. It is reasonably common to implement a programming language using a virtual machine;

3.2 NET BEANS

Net Beans IDE is the official IDE for Java 8. With its editors, code analyzers, and converters, you can quickly and smoothly upgrade your applications to use new Java 8 language constructs, such as lambdas, functional operations, and method references. Batch analyzers and converters are provided to search through multiple applications at the same time, matching patterns for conversion to new Java 8 language constructs. With its constantly improving Java Editor, many rich features and an extensive range of tools, templates and samples, Net Beans IDE sets the standard for developing with cutting edge technologies out of the box.

3.3 ECLIPSE

In computer programming, Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Written mostly in Java, Eclipse can be used to develop applications. By means of various plug-ins, Eclipse may also be used to develop applications in other programming languages: Ada, ABAP, C, C++, COBOL, FORTRAN, Haskell, JavaScript, Lasso, Natural, Perl, PHP, Prolog, Python, R, Ruby (including Rubyon Rails framework), Scala, Closure, Groovy, Scheme, and Erlang. It can also be used to develop packages for the software Mathematica. Development environments include the Eclipse Java development tools (JDT) for Java and Scala, Eclipse CDT for C/C++ and Eclipse PDT for PHP, among others. The initial codebase originated from IBM Visual Age. The Eclipse software development kit (SDK), which includes the Java development tools, is meant for Java developers. Users can extend its abilities by installing plug-ins written for the Eclipse Platform, such as development toolkits for other programming languages, and can write and contribute their own plug-in modules. Released under the terms of the Eclipse Public License, Eclipse SDK is free and open source software (although it is incompatible with the GNU General Public License). It was one of the first IDEs to run under GNU Class path and it runs without problems under Iced Tea.

3.4 DROP BOX

Drop box is a file hosting service operated by Drop box, Inc., headquartered in San Francisco, California, that offers cloud storage, file, personal cloud, and client software. Drop box allows users to create a special folder on their computers, which Drop box then synchronizes so that it appears to be the same folder (with the same contents) regardless of which computer is used to view it. Files placed in this folder are accessible via the folder, or through the Drop Box website and a mobile app. Drop box was founded in 2007 by Drew Houston and Arash Ferdowsi, as a Y Combinatory startup company.

III. RESULTS

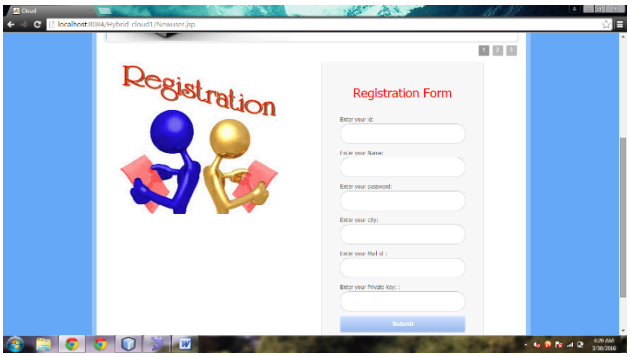

Fig 2. Registration

Fig 3. File Upload

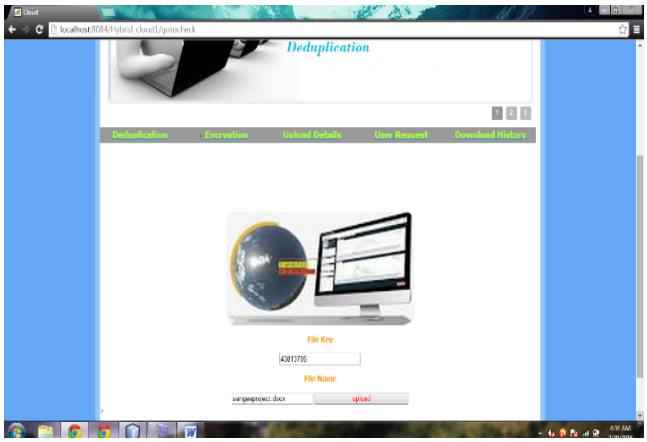

Fig 4. Key Request

Fig 5. Third party access

Fig 6. File Download

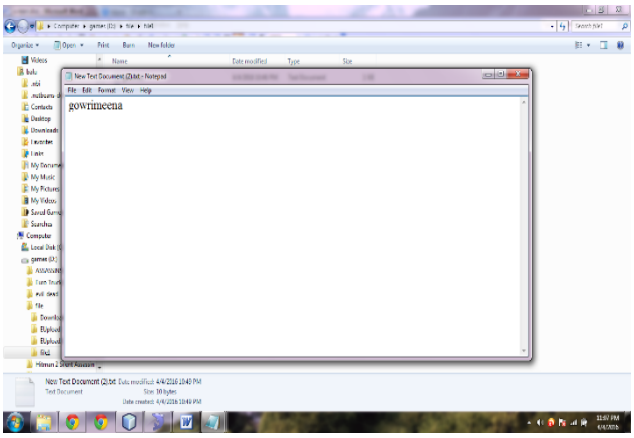

Fig 7. Decrypt file

IV. CONCLUSION AND FUTURE ENHANCEMENT

1. Conclusion

In this paper, the notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. In presented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct test bed experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

REFERENCES

[1] OpenSSL Project. http://www.openssl.org/.

[2] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart. Duplass: Serve raided encryption for reduplicated storage. In USENIX Security Symposium, 2013

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT,pages 296– 312, 2013.

[5] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.

[6] M. Bellare and A. Palacio. GQ and Schnarr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002.

[7] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011

[8] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, “Reclaiming space from duplicate files in a server less distributed file system,” in Proc. Int. Conf. Disturb. Compute. Syst., 2002, pp. 617–624.

[9] D. Ferraiolo and R. Kuhn, “Role-based access controls,” in Proc. 15th NIST-NCSC Nat. Compute. Security Conf., 1992, pp. 554–563.

[10] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Proofs of ownership in remote storage systems,” in Proc. ACM Conf. Compute. Common. Security, 2011, pp. 491–500.

[11] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," in Proc. IEEE Trans. Parallel Disturb.

[12] C. Ng and P. Lee, "Revdedup: A reverse deduplication storage system optimized for reads to latest backups," in Proc. 4th Asia- Pacific Workshop Syst., http://doi.acm.org/10.1145/2500727. 2500731, Apr. 2013.

[13] W.K.Ng, Y.Wen, and H.Zhu, "Private data deduplication protocols in cloud storage," in Proc. 27th Annu. ACM Symp. 2012, pp. 441-446.