



Combining Deduplication and String Comparison for Avoiding Redundant Data with Enhanced Authentication Approach on Cloud

R.Sasikumar¹, S.Deva Priya²

Department of Computer Science & Engineering, K.Ramakrishnan College of Engineering, Samayapuram, Trichy¹

Department of Computer Science & Engineering, SRM TRP Engineering College, Irungalur, Trichy²

Abstract: Every day activity many number of data and file has been generated. Every organization wants to store and main that information with efficient manner. Efficient denotes, storage space utilization, securely maintaining the organization records, accessing records quickly. On the organization server maintain the files some of them are stored for more than once. It will increase the utilization of memory. It will lead us to insufficient storage space. Sometimes outgoing data in the network connection contain the duplicate files so, network traffic is occurring (or) the transfer speed gets shrunk. So, here we propose a technique called data deduplication and we also use the algorithm of machine learning called string comparison to detect the redundant data and files. To ensure the security we also proposed hybrid authentication approach.

Keywords: Cloud Services, Deduplication, Greedy Approach, String Comparison

I. INTRODUCTION

Consistently on the planet delivered an enormous measure of documents and information. The client can store their record and information in local PC, remote server, database, and cloud. Distributed computing is the usage of remote servers on the web to store, oversee and process information as opposed to a local server or personal computer. The motivation behind Cloud is utilized to store and process a record and data which is transferred by clients and association. There are part cloud suppliers who give the cloud administration to the clients who need an asset, however not have the options to manage the cost of more cash to store and process. The cloud gives three different ways of administrations to the clients in particular IAAS (Infrastructure as a Service), PAAS (Platform as a Service), SAAS (Software as a Service).

Software as a Service (SaaS), speak to the biggest cloud showcase are as yet developing rapidly. SaaS utilizes the web to convey applications that are managed by a third-party and whose interface is gotten to on the customers' side. Most SaaS applications can be run straightforwardly from an internet browser with no downloads or establishments required, although some additional plugins. Platform as a Service (PaaS), are used for applications, and another improvement when giving cloud services to software. Designers can gain with PaaS is a system they can expand upon to create or modify applications. PaaS makes the advancement, testing, and deployment of application fastly, straightforward, and savvy. With this innovation, venture activities, or an outsider supplier, can oversee Operating Systems, virtualization, servers, storage, organizing, and the PaaS architecture itself. Designers, in any case, deal with the applications.

Infrastructure as a Service (IaaS) provides virtualization of computing resources through internet. The above mentioned services can be utilized any of the following categories: Public cloud, Private cloud, and Hybrid cloud. In public cloud computing resources like Servers, Storage, Networks and development platform can shared among multiple companies. Also it supports any one can access these resources those who subscribed the cloud. The issues we see with the open cloud—low visibility into traffic and action, security concerns. Because of the above mentioned problems we are going to concentrate only Private cloud. The individual endeavour made and kept up a private cloud. Private clouds are committed to a solitary association or business. It underpins a group of the devoted client not for everybody in the network. Along these lines, it is appropriate to store confidential information of the association. While we are sparing our documents on the cloud we can lessen the measure of extra space and the bandwidth capacity. It makes the framework increasingly secure, the various benefits of clients are again considered while checking duplicate content. Be that as it may, the issue which happens with this methodology is that regardless of whether including the additional substance in the records and put away the information in the cloud. So as to validate the client information prompts a decrease in distributed storage space.



The solution to this problem is to perform Verification and filtering process. In this approach, the file must be stored in a safe and secure manner. If there is any intruder in our organization who goes to make illegal access to our file and make a modification[9]. We can easily identify and block them from login again and also restrict the access rights of that user. Hence, users can keep their file safe. The Deduplication technique used to eliminate duplicate files from both server and cloud. The string comparison is an algorithm which is used to compare the files to identify the modified or unmodified duplicate file and also used the filter algorithm to identify which file is duplicated. String comparison and filtering algorithm both are coming under the machine learning technique. Normally machine learning is used for in depth analysis of data.

II. RELATED WORKS

Considering data duplication and authentication of confidential information outsource to a cloud, paper has been proposed. For secure storage the system by Jiawei Yuan and Shucheng Yu based on Secure[11] and constant public cloud storage for deduplication. Data Deduplication process helps to reduce the data and file redundant. This makes the efficient storage and bandwidth utilization.

Data Integrity and Security are two major concerns on the cloud platform. By implementing Data Deduplication and String comparison redundant data can be reduce[12]. With help of above techniques integrity can implement. In order to achieve efficient storage allocation and accessing we proposed Greedy based technique called Huffman Encoding. With help of above technique we can achieve efficient file access. Existing system does not provide storage efficiency.

III. PROPOSED SYSTEM

The proposed system uses Data Deduplication and String comparison concepts to avoid redundant data and file duplication. And the Hybrid authentication mechanism helps to verify the authentication of user. Greedy approach also used to compress data to provided better storage utilization.

The following sections will describe the overall proposed architecture and the concepts of Data Deduplication, Optimization, String Comparison and Filtering, Authentication mechanisms [13] used.

3.1 Intelligent compression of Data Deduplication

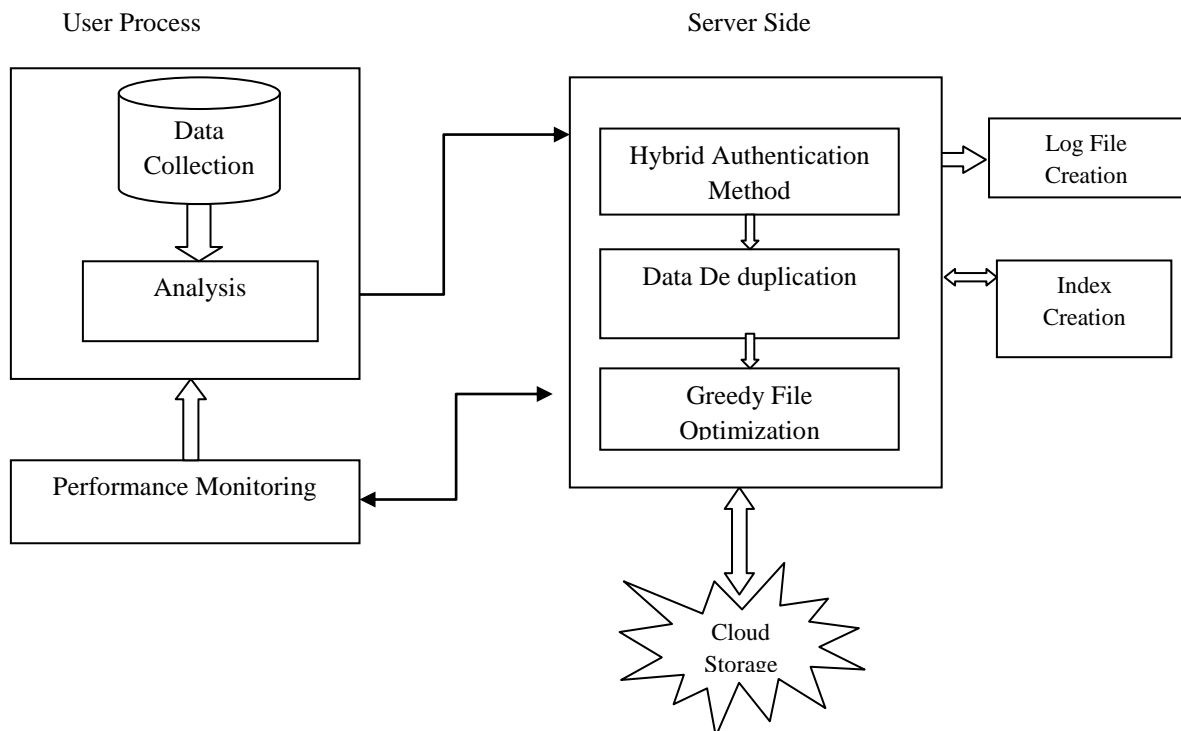


Fig 3.1: Proposed System Architecture

This process eliminates the duplication of data and redundant instance of file. It also called as Singles instance data. It is implemented based on the mechanism called data backup and network data. It increases the storage of unique data within a cloud. This strategy has worked by examining and contrasting each incoming information and the information which exist in our storage. On the off chance that information is as of now present in the database this system dispenses

with the new information and simply makes a reference for that new approaching information. In the event that there are any changes in the approaching information, at that point that changes are just refreshed with the current document then recently coming record is erased; however, there is dependably a reference for every approaching document. Reference is nothing but the information about which file is newly entered in to system.

3.2 Huffman Encoding for File Optimization

The proposed method consists of Greedy Approach to store the file in the cloud. It provides best optimization solution in terms of compressing data. The major demerits of this algorithm is software attacks occurred because of mobile malware such as Trojans, Worms and Viruses can result in privacy leakage, economic loss, power depletion, and network performance degradation of the system.

3.3 Proposed Client Authentication in the cloud

Authentication mechanism gives access control to frameworks by verifying whether a client's credential matches the accreditations in a database of approved clients or in an information verification server[1]. Once the credentials matches with existing credentials it allows user to access their respective files. Otherwise it stores the information of system where the unauthorized access origins. This information helps to prevent from the future access fro the same machine or user.

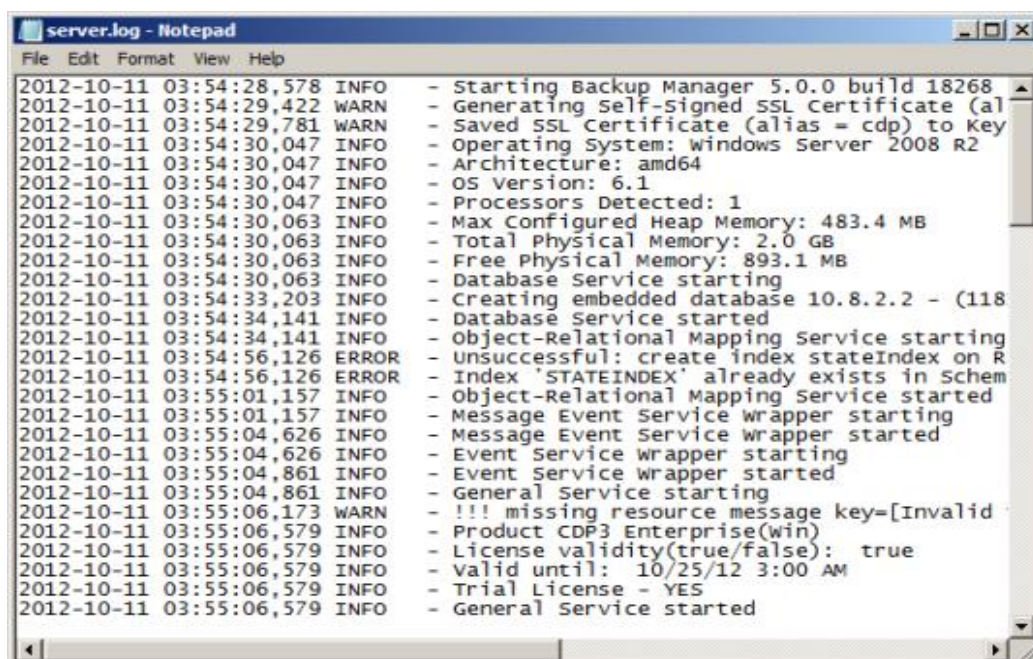
Here we propose Hybrid Authentication mechanism. It comprises of both One-Time Password (OTP) Authentication and HTTP based API authentication.

- **One-Time Password Authentication:[8]** It is automatically generated secret code based on Random Key Generation method. It can be combinations of numeric and alphabets. This code is valid for 15 minutes from generation.
- **HTTP based API Authentication:** Here server requests the client authentication in the form of user name and password through the browser.

3.4 String Comparison and Filter Algorithm

When the user need to ensure there is no duplication. User uses the string comparison it will compare the file in the local server and cloud with the same file in the database. The algorithm first convert the data into binary format then compared with data in the database. Because the data in database is in binary format we cannot convert that in to normal format to compare. If the string mismatch will occur at any place while comparing the file. Then there is duplication on that file which means someone accesses the file and made a modification on that illegally. Filter Algorithm is an approach to improving quality of raw data collected from various sources. It is most effective in cases when there is in band noise present. This algorithm helps us to identify which file is modified by using the generated log file. The log file contains which file is modified, when that file modified and how much data is modified. The file stored in the local server only gets changed. The backup copy of the file in the database cannot change.

3.5 Log file generation



```

server.log - Notepad
File Edit Format View Help
2012-10-11 03:54:28,578 INFO - Starting Backup Manager 5.0.0 build 18268
2012-10-11 03:54:29,422 WARN - Generating Self-signed SSL certificate (a)
2012-10-11 03:54:29,781 WARN - Saved SSL Certificate (alias = cdp) to key
2012-10-11 03:54:30,047 INFO - Operating System: windows server 2008 R2
2012-10-11 03:54:30,047 INFO - Architecture: amd64
2012-10-11 03:54:30,047 INFO - OS version: 6.1
2012-10-11 03:54:30,047 INFO - Processors Detected: 1
2012-10-11 03:54:30,063 INFO - Max Configured Heap Memory: 483.4 MB
2012-10-11 03:54:30,063 INFO - Total Physical Memory: 2.0 GB
2012-10-11 03:54:30,063 INFO - Free Physical Memory: 893.1 MB
2012-10-11 03:54:30,063 INFO - Database Service starting
2012-10-11 03:54:33,203 INFO - Creating embedded database 10.8.2.2 - (118
2012-10-11 03:54:34,141 INFO - Database Service started
2012-10-11 03:54:34,141 INFO - Object-Relational Mapping Service starting
2012-10-11 03:54:56,126 ERROR - unsuccessful: create index stateIndex on R
2012-10-11 03:54:56,126 ERROR - Index 'STATEINDEX' already exists in schem
2012-10-11 03:55:01,157 INFO - Object-Relational Mapping service started
2012-10-11 03:55:01,157 INFO - Message Event Service wrapper starting
2012-10-11 03:55:04,626 INFO - Message Event Service wrapper started
2012-10-11 03:55:04,626 INFO - Event Service wrapper starting
2012-10-11 03:55:04,861 INFO - Event Service wrapper started
2012-10-11 03:55:04,861 INFO - General service starting
2012-10-11 03:55:06,173 WARN - !!! missing resource message key=[Invalid
2012-10-11 03:55:06,579 INFO - Product CDP3 Enterprise(win)
2012-10-11 03:55:06,579 INFO - License validity(true/false): true
2012-10-11 03:55:06,579 INFO - Valid until: 10/25/12 3:00 AM
2012-10-11 03:55:06,579 INFO - Trial License - YES
2012-10-11 03:55:06,579 INFO - General Service started

```

Fig. 3.2: Log File



When the duplication and duplicated file is identified then log file is generated on the cloud. At the same time only file in the local server get changed. The file copy in the database is never get modify. If intruder modify the file directly this is in cloud also identified. By using the generated log file in the cloud we can identify the modified file name when the file modified (date and time). The following output screen shows the sample log details. It shows the details of duplication files with Time stamp and information about particular file.

IV. CONCLUSION

In this paper, we propose a method called Deduplication and String Comparison to avoid redundant data and file. In cloud environment, we can provide secure access control by implementing hybrid authentication method. In addition to authentication, we are using Filter methods to avoid unauthorized access through the help of log file. In future enhancements, we have planned to implement novel cryptography algorithm to provide enhanced security. A greedy algorithm always generated locally optimal solutions. The major demerits of this algorithm is software attacks occurred because of mobile malware such as Trojans, Worms and Viruses can result in privacy leakage, economic loss, power depletion, and network performance degradation of the system. So further improvement is necessary to implement in a secure way.

REFERENCES

- [1]. T. Dillon, C. Wu, and E. Chang, "Cloud Computing: Issues and Challenges," 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 27-33, DOI= 20-23 April 2010.
- [2]. J. F. Yang and Z. B. Chen, "Cloud Computing Research and Security Issues," 2010 IEEE International Conference on Computational Intelligence and Software Engineering (CiSE), Wuhan pp. 1-3, DOI= 10-12 Dec. 2010.
- [3]. M. M. Alabbadi, "Cloud Computing for Education and Learning: Education and Learning as a Service (ELaaS)," 2011 14th International Conference on Interactive Collaborative Learning (ICL), pp. 589 – 594, DOI=21-23 Sept. 2011.
- [4]. C.S Lee, G.M Lee, W.S Rhee, "Smart Ubiquitous Networks for future telecommunication environments", Computer Standards & Interfaces, Vol. 36, pp. 412–422, 2014.
- [5]. Anuj Pathania, Vanchinathan enktramani, Muhammad Shafique, Tulika Mitra, and Jörg Henkel "Optimal Greedy Algorithm for Many-Core Scheduling" iee transactions on computer-aided design of integrated circuits and systems, VOL. 36, NO. 6, JUNE 2017.
- [6]. Yen-Ching Hsu, Kuan-Li Peng, Chin-Yu Huang, "A study of applying severity-weighted greedy algorithm to software test case prioritization during testing", IEEE International Conference on Industrial Engineering and Engineering Management, 2014.
- [7]. References from <https://www.techopedia.com/definition/16931/greedy-algorithm>
- [8]. Referred from the site "<https://searchsecurity.techtarget.com/definition/authentication>".
- [9]. Christin, Nicolas & Safavi-Naini, Reihaneh. (2014). Financial Cryptography and Data Security: 18th International Conference, FC 2014, Christ Church, Barbados, March 3-7, 2014, Revised Selected Papers. 10.1007/978-3-662-45472-5.
- [10]. Jiawei Yuan & Shucheng Yu(2013) 'Secure & constant cost public cloud storage auditing with deduplication' IACR cryptology ePrint archive.
- [11]. B.Kiran Bala , J.Lourdu Joanna, Multi Modal Biometrics using Cryptographic Algorithm, European Jour of Academic Essays 1(1): 6-10, 2014.
- [12]. B.Kiran Bala, A Novel Approach to Generate a Key for Cryptographic Algorithm, Journal of Chemical and Pharmaceutical Sciences, (JCHPS), Special Issue 2: February 2017, Page 229-231.