Securing Data Duplication: A Method for Eliminating Duplicate Records

Mr. G Ahmed Zeeshan[1], Dr. Y Ravi Kumar[2], Dr. E Mohan[3]

Assistant Professor[1], Associate Professor[2], Professor [3]
Department of CSE
Global Institute of Engineering and Technology

ABSTRACT

We get a hash value using MD5 hashing whenever a file is uploaded to cloud storage. To find duplicate data in the cloud, we compare those values to those acquired when the identical item is uploaded under a different name. A plan for the safe modification of data throughout the organization when deduplication is finished. Security is accomplished via the use of decryption and encryption algorithms. Specifically, this study looks at the secure deduplication method. Following the removal of duplicate content, pointers will make reference to the original file.

1. INTRODUCTION

Cloud computing makes use of several techniques in Platform as a Service (PaaS), an excellent platform for the developer to create web applications. As a service to the requester, the IaaS processing framework may be sent out. Regarding the Virtual Machine (VM) component of your current application architecture.

When it comes to determining your present strong performance, cloud planning plays a crucial role, and cloud computing is still in its infancy with numerous concerns and obstacles. The use of the cloud for online storage and application development is growing rapidly. Cost and on-interest administrations are two areas where cloud computing offers some benefits. Distributed computing based on real-time communication, such as PC got new ways of thinking. Because of capacity and organization, cloud computing now stores the vast majority of the world's data. Copy information shows up on the information plate, but the plate can't read it. Details about the duplicate might affect the available space in the circle. When data is stored and addressed using a conventional technique, copy information becomes visible. Finding instances of duplicate data is a pain. Organizational information and unstructured information are the two main categories of data that are playing a major role in the current trend. Site logs, customer call detail records, and other similar types of information are usually a part of the well-coordinated construction data.

Blog posts, media collaboration details, audio recordings, and other forms of unstructured data are becoming more difficult to manage as a result of the explosion in mobile media consumption and online video. This means that unstructured data has to be realistically monitored. It now accounts for almost 13% of IT budgets allocated to capacity1. More problems arise as a result of these consequences, such as a decline in execution quality, a loss in bargain value, and higher operating costs. The goal, therefore, is to overcome these problems and deal with the framework where the concept of deduplication is established. Both the square level (subrecord) and the document level are investigated by deduplication innovation. Smaller, fixed or variable squares or parts make up the approaching data.

A unique identity is generated for each of these smaller squares using a series of hashing algorithms or even a little correlation of the square. For this cycle, MD5 is the standard for computations. As a further resource, content aware reasoning considers the information's substance type when drawing conclusions about its quantity and limitations.

2.LITERATURE SURVEY

An examination of the risks associated with cloud computing

Although cloud computing has been around for a while, it has only recently grown in popularity. As long as there is an internet connection, consumers may conduct computer tasks anytime and anywhere using this pay-per-use service. Public, private, community, and hybrid cloud deployment strategies are the four most common types. The significant benefits of the cloud are the source of its prominence. Nevertheless, the widespread use of this technology might be impeded by security concerns over the cloud's capacity to protect, maintain, and audit data. Therefore, it is really a crucial step to decide whether to use on-premises IT resources or switch to a cloud service provider. This article gives a general outline and detailed description of cloud computing. In addition to outlining the suggested security challenges and their possible solutions, it compiles a list of the most pressing security threats facing cloud providers and their customers today.

Technical Education Community Leadership

Professional learning communities (PLCs) have a favorable effect on both teaching methods and student achievement, according to an analysis of PLC features. Only a few of empirical research have looked at how it affects classroom instruction and student achievement.Hence, the practice of teaching and learning is being worked on. Therefore, we are launching a new community to facilitate online education, instruction, and the dissemination of empirical study results. The community's combined findings from these investigations will imply that pedagogical practices and student outcomes are interrelated. This project's implications and recommendations for further research on PLCs' effects on classroom instruction are detailed here.

Examining ETS via Android Devices:

Our daily lives are being profoundly affected by the exponential expansion of android apps. In order to automate the process of employee monitoring in the firm via their workplace cell phone and to boost organizational development, this survey is employing an android mobile system. Using Android technology, this paper details the design and implementation of an admin app, an employee app, and a centralised server for monitoring corporate employees. A dynamic database utility is available in this system to access information stored in a central database. Every aspect of an employee's phone usage, including their location, web browser history, data uses, illegal data uses, and employee SMS history, is stored in the android app on smart phones. The employee's phone and the administrator's are always in constant contact thanks to the 3G network. The interface of this program is intuitive. In addition to saving time and decreasing management effort, this approach also enhances accuracy in controlling the firm's personnel by preventing them from using corporate phones for anything other than work. Additionally, this system may access the whole history of employee phone usage via its connection to the centralized server. Managers can know the good, the average, and the bad conduct of their employees by navigating them via mobile phones, which is the key component of our study.

Reconciling end-to-end secrecy and data minimization in cloud storage.

When storing data on external storage devices or in the cloud, end-to-end encryption is becoming standard practice among storage system users. Unfortunately, this approach hinders the advantages of compression and deduplication that are carried out after data encryption, which leads to a rise in both the storage space needed and the total cost of the service.

In this research, we tackle this issue by presenting a system that integrates end-to-end encryption with compression and deduplication in the downstream processes. Even if customers terminate their cloud storage subscription, the suggested architecture ensures data secrecy both while in transit and at rest, so long as storage systems can continue to execute data reduction activities. The framework doesn't need any changes to the client's business applications and only necessitates small tweaks to data encryption storage apps. We also provide many safe data reduction techniques that can deduplicate and compress data without eavesdropping, regardless of whether key was used to encrypt it initially. We provide a thorough security study that proves the framework is safe against hostile cloud administrators, renters, and authorities. The architecture allows for significant storage capacity reductions, as shown by our prototype, for a fair additional cost in the time needed to store data.

3. EXISTING SYSTEM

To study the data breaches in cloud storage, a study was carried by. Various instances of breaches were found where the data of the client was exposed by the service providers. The instances exposed that if the service provider or the client has access to data of other users the breaching of data was more. For handling the data breach problem, the authors suggested end-to-end encryption. The issues in deduplication while encryption were found by authors in. To resolve they proposed a novel encryption methodology. In the methodology, the encryption units were transformed into chunks and these chunks were used to produce symmetric keys. The symmetric key obtained were used to limit mapping between plain and ciphertext. To reclaim space that was lost during replicating files, a methodology was introduced by. The methodology involved convergent encryption that permitted duplicate files to be consolidated into a single file using diverse user keys and SALAD, a Self Arranging Lossy Association Database.

The authors in proposed FadeVersion, a system for cloud backup which also can act as a security layer. It was also able to provide cryptographic security to data.

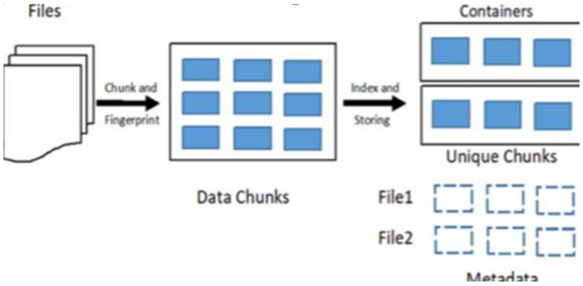
1.2 PROPOSEDSYSTEM

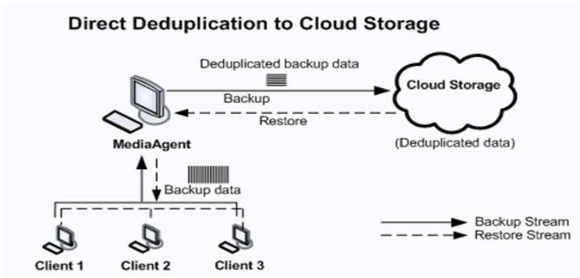
Presently day information duplication is a quickly developing method used in information reinforcement stored without excess. It is vital special and unique. In this paper, we design an interactive protocol using AWS services in which we use the AWS lambda function for generating the hash value of the file which gets uploaded. We use AWS cloud watch for records of every file, also used the S3 bucket for storing and retrieving the data. We investigated the information to decide the relative adequacy of information deduplication, especially considering the entire record versus the block-level end of excess. Security in information deduplication can be furnished with the utilization of a concurrent encryption method that encodes the information previously transferred to the public framework.

To prove the thought, we proposed the model and attempted some tests, in that test we uploaded the same files with different names and different file systems such as pdf, doc, odt. The work shows that the proposed system works correctly and gives a warning that the same file was uploaded before. Cloud computing is productive and adaptable yet keeping up the strength of preparing such countless positions in the distributed computing climate the cloud framework faces the issues of replication furthermore, the information duplication as indicated by situations.

In this setting need to tackle the issue of both, to upgrade the cloud execution as far as capacity overhead and accessibility needed to deal with the whole information in such a way by which the hunting capacity and the ordering of information can be accomplished both..

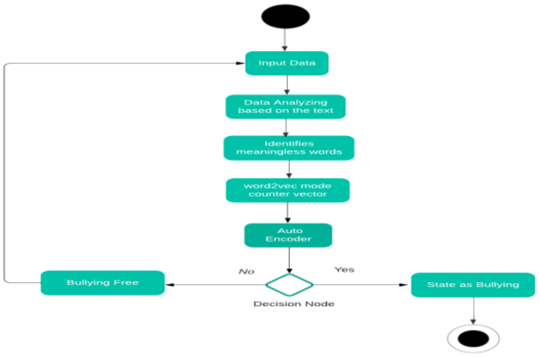
4.SYSTEM ARCHITECTURE .





Activity Diagram

A graphical representation of the work process of stepwise exercises and activities with support for decision, emphasis and simultaneousness, used to depict the business and operational well-ordered stream of parts in a framework furthermore demonstrates the general stream of control.


Input Data
Data Analyzing based on the text
Identifies meaningless words
word2vec mode counter vector
Auto Encoder
No
Yes
Bullying Free
Decision Node
State as Bullying

5. SYSTEM IMPLEMENTATION

MODULES

There are 2 modules:

1. User
2. Cloud

User: -

- Register
- Login
- Data Storage
- Data search
- Profiles
- Downloads Files
- Logout

Cloud: -

- Login
- Manage Users
- View files
- User Authentication

6.1 TYPES OF TESTING

■Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

■Integration testing

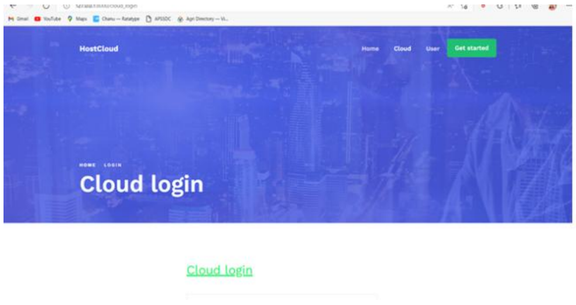
Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

■Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

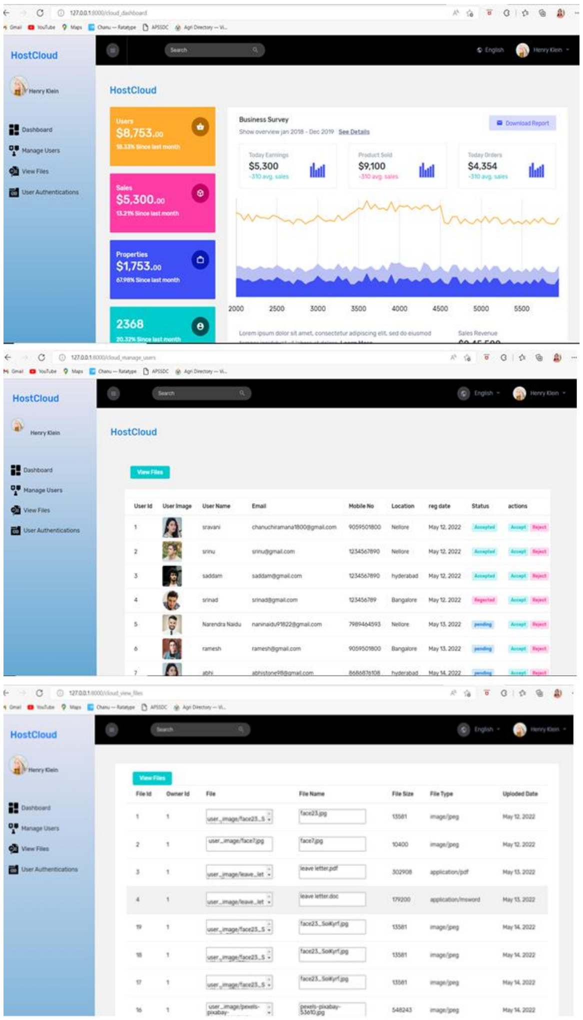
7.RESULTS



Cloud login

Cloud login

fig.7. 1 Home page



User Id	User Image	User Name	Email	Mobile No	Location	reg date	Status	actions
1		sravani	chanuchiramana1800@gmail.com	9059501800	Nellore	May 12, 2022	Accepted	Accept Reject
2		srinu	srinu@gmail.com	1234567890	Nellore	May 12, 2022	Accepted	Accept Reject
3		saddam	saddam@gmail.com	1234567890	hyderabad	May 12, 2022	Accepted	Accept Reject
4		srinad	srinad@gmail.com	123456789	Bangalore	May 12, 2022	Rejected	Accept Reject
5		Narendra Naidu	naninaidu91822@gmail.com	7989464593	Nellore	May 13, 2022	pending	Accept Reject
6		ramesh	ramesh@gmail.com	9059501800	Bangalore	May 13, 2022	pending	Accept Reject
7		abhi	abhistone98@gmail.com	8686876108	hyderabad	May 14, 2022	pending	Accept Reject

File Id	Owner Id	File	File Name	File Size	File Type	Uploded Date
1	1	user_image/face23_S	face23.jpg	13581	image/jpeg	May 12, 2022
2	1	user_image/face7.jpg	face7.jpg	10400	image/jpeg	May 12, 2022
3	1	user_image/leave_let	leave letter.pdf	302908	application/pdf	May 13, 2022
4	1	user_image/leave_let	leave letter.doc	179200	application/msword	May 13, 2022
19	1	user_image/face23_S	face23_SoiKyrf.jpg	13581	image/jpeg	May 14, 2022
18	1	user_image/face23_S	face23_SoiKyrf.jpg	13581	image/jpeg	May 14, 2022
17	1	user_image/face23_S	face23_SoiKyrf.jpg	13581	image/jpeg	May 14, 2022
16	1	user_image/pexels-pixabay-	pexels-pixabay-53610.jpg	548243	image/jpeg	May 14, 2022

8. CONCLUSION & FUTURE WORK

If you are a new examiner and need to look into safe data deduplication, this article is for you. We want to enhance the implementation of our planned work in security outlook by using the security methods that have been accumulated here. Deduplication may be used with mixed data thanks to a basic approach. In order to duplicate data and send it securely via a distributed processing environment, a mechanism must be developed. We are developing a deduplication framework that makes use of one of the MD5 hashing algorithms, as well as an extra security measure for safe data transfer that makes use of AWS's features.

REFERENCES

In 2013, Bhoyar and Chopde published a study. "Cloud computing: Service models, types, database, and issues"Third issue of the International Journal of Advanced Research in Computer Science and Software Engineering.
Referenced in [2] Kaur and Singh (2015). Critical challenges with cloud computing security reviewed.Vol. 8, No. 3, pp. 397, International Journal of Engineering and Technology Advancements.
Referenced in Pathan (2017). Idea: Community Management for Tech-Based Learning. Volume 5, Issue 1631 of the International Journal of Scientific Research and Development (IJSRD).

The authors of this work are Pathan and Shaikh (2018). A Survey on ETS Using Android Phone. The citation is from the International Journal of Innovative Research in Technology (IJIRT), volume 5, issue 3, page.

The authors of the cited article are Baracaldo, Sorniotti, Glider, and Androulaki (2014, November). Reconciling end-to-end secrecy and data minimization in cloud storage.Presented at the ACM Workshop on Cloud Computing Security, Sixth Edition, (pp. 21–33).

The authors of the cited work are Wang, C., Qin, Z. G., Peng, J., and Wang, J. (2010, July). An innovative encryption method for data deduplication technology. On pages 265-269, at 2010's International Conference on Communications, Circuits and Systems (ICCCAS).IEEE.

In a 2002 July publication, Douceur, Adya, Bolosky, Simon, and Theimer were listed as authors. Getting rid of duplicate files in a distributed file system that doesn't need servers. Volume 22, Issue 6, Pages 617-624, Proceedings of the 22nd International Conference on Distributed Computing Systems. IEEE.

In a September 2011 publication, Rahumed et al. [8] discussed the work of Chen, Tang, Lee, and Lui. An guaranteed deletion and version control cloud backup solution that is both safe and convenient.Found on pages 160–167 of the 2011 40th International Conference on Parallel Processing Workshops.IEEE.