FINDING DATA DEDUPLICATION USING CLOUD

Mr. M. A. R. KUMAR, Mrs. SRILATHA PULI,
Associate Professor in Sreyas Institute of Engineering and Technology, JNTUH, India,anandranjit@sreyas.ac.in
Assistant Professor in Sreyas Institute of Engineering and Technology, JNTUH, India,srilatha.puli@sreyas.ac.in
TEAM MEMBERS:
YANAGANDULA NANDINI-18VE1A0558- nandiniyanagandula@gmail.com
K. PRASANNA VAISHNAVI-19VE5A0504-prasannavaishnavi20@gmail.com
NAREDLA MEGHANA-18VE1A0535- naredlameghana2000@gmail.com
RUDAVATH SAI -18VE1A0545 - rudavathsai0@gmail.com

ABSTRACT-

Data grows at the emotional rate of 50% per time, and 75% of the digital world is a copy1 Although keeping multiple clones of data is necessary to guarantee their availability and high continuity and the quantum of data redundancy is inordinate. By keeping a single dupe of repeated data, data deduplication is one of the most promising results to reduce the storage costs, and improve users experience by saving network bandwidth and reducing provisory time. However, this result must now solve many security issues to be fully satisfying. In this project we target the attacks from malicious clients that are grounded on the manipulation of data identifiers and those based on backup time and network traffic observation. Our system provides global storage space savings, per-customer bandwidth network savings between clients and deduplication proxies, and saving global network bandwidth between deduplication proxies and the storage server. The evaluation of our result compared to a classic system shows that the overhead introduced by our scheme is mostly due to data encryption which is necessary to ensure data confidentiality. Data deduplication allows the cloud users to manage their cloud storage space for storing effectively by avoiding storage of repeated data's and save bandwidth. Here we use the Cloud Me for the data storage. For data confidentiality the data are stored in an encrypted form using Advanced Encryption Standard (AES) algorithm.

KEYWORDS *– data de-duplication, cloud computing, encryption, data confidentiality*

1.Introduction

The amount of knowledge to be stored by cloud storage systems increases extremely fast. It is thus of utmost importance for Cloud Storage Providers (CSPs) to dramatically reduce the value to store all the created data. A promising approach to realize this objective is through data deduplication. Data deduplication keeps a single copy of redundant data. When a client wants to store some amount of data, and if a copy of this data has already been saved in the storage system, then the reference to this existing copy is stored at the storage server. There is no duplication is created.

There are various forms of data deduplication. It can be done by a client directly on the data he/she has previously stored in the system, a technique commonly called intra-user deduplication, or it can be achieved by taking into account the data previously stored by all the clients.

In this case it is taken as inter-user deduplication. Data deduplication improves users experience by saving network bandwidth and saving backup time when the clients perform the duplication before uploading data to the server. This form of deduplication is termed as client-side deduplication, and when it is handled by the storage server it is called server-side deduplication. Data deduplication is gaining popularity in both commercial and research storage systems.

Therefore, many works have recently revealed some major security issues leading to information leakage to malicious clients. These security issues arise mainly in systems performing an inter-user and client-side deduplication which is unfortunately this deduplication provides the best savings in terms of network bandwidth and storage space.

2.LITERATURE SURVEY

This case study describes the data deduplication and other methods of reducing storage consumption play an important role in affordably managing today's explosive growth of data. Reducing the use of storage is part of a broader strategy to provide an efficient information infrastructure that is responsive to dynamic business requirements. This will explore the significance of deduplication ratios related to particular capacity optimization techniques within the context of information lifecycle management.

3.EXISTING SYSTEM

Data deduplication keeps a single copy of redundant data. When a client wishes to store some small amount of data, and if a copy of this data has already been saved in the storage system, then a reference to this existing copy is stored at the storage server. There is no duplication is created. There are various forms of data deduplication. It can be done by a client alone on the data he/she has previously stored in the system, a technique commonly called intra-user deduplication, or it can be achieved by taking into account the data previously stored by all the clients. In this case it is referred as inter-user deduplication. Data deduplication also improves users experience by saving network bandwidth and backup time when the clients perform the data deduplication before uploading data to the storage server. This form of deduplication is termed as client-side deduplication, and when it is handled by the storage server it is called server-side

deduplication. Due to its straightforward economical advantages, data deduplication is gaining popularity in both commercial and research storage systems. Many works have recently revealed major security issues leading to information leakage to malicious clients. These security concerns arise mainly in systems performing an inter-user and client-side deduplication which is unfortunately this deduplication provides the best savings in terms of storage space.

DISADVANTAGES

- Duplication of data in separate files when data is updated
- Lack of data integrity
- A situation in which program and data organized for one application are incompatible another application.

4. PROPOSED SYSTEM

In finding data duplication using cloud we save storage space and network bandwidth as it masters all the network and storage infrastructure, and provide a secure storage service to its consumers.

Our deduplication scheme is simple and robust and is efficient in terms of storage space and bandwidth savings for both clients and cloud service provider. We consider data deduplication at a file level granularity but our solution can be extended to the block level. Here our approach is a two-phase deduplication that combines both intra- and inter-user deduplication techniques by introducing deduplication proxies between the clients and the storage server. Communications from clients go through these DPs to reach the SS which allows splitting the deduplication process.

ADVANTAGES

- This offers more security to the sensitive data in the proposed security model.
- Reduce the unwanted storage space utilization due to data redundancy.
- To improve the security and protect the data confidentiality.
- The user is only allowed to perform duplicate check for files marked with
  Corresponding privileges.
- Deduplication lowers storage costs as fewer disks are needed.

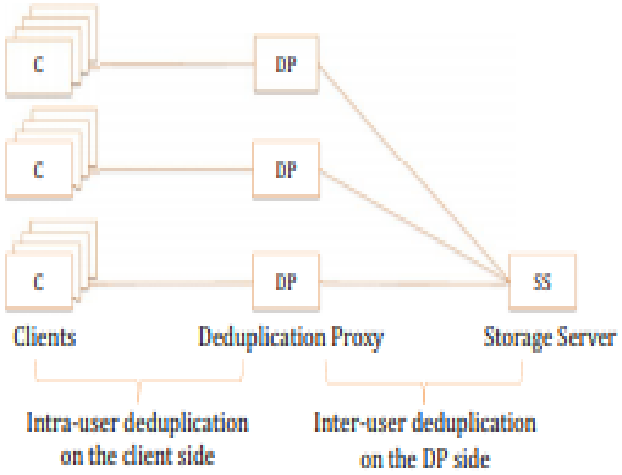

Figure 1: Architecture Diagram for Finding Data Duplication Using Cloud

5. IMPLEMENTATION

Implementation includes all those activities that covert from old system to new system. The old system consists of manual operations, which is operated in a very difficult manner from the proposed system. A proper implementation is essential to provide a reliable system to meet the requirements of the organization.

Data Owner:
Data owner can outsource a collection of encrypted files to Cloud, he can perform add, delete, user request processing operation on files.

Data User:
Data user can perform keyword searches on the cloud to get desired documents, and he can send the access request to data owner to get decrypted file.

Storage Server (SS):
A server in charge of storing and serving clients files. The storage server also maintains an index of all the files stored in the storage system and their owners.

Deduplication Proxy (DP):
A server consists of given number of clients. Clients communicate with the Storage Server via their associated deduplication proxy. A deduplication proxy is involved in both the intra-user and the inter-user deduplication.

Functional Requirements:
Functional requirement should include function performed by a specific screen outline work-flows performed by the system and other business or compliance requirement the system must meet.
Functional requirements specify which output file should be produced from the given file they describe the relationship between the input and output of the system, for each functional requirement a detailed description of all data inputs and their source and the range of valid inputs must be specified.
The functional specification describes what the system must do, how the system does it is described in the design specification.
If a user requirement specification was written, all requirements outlined in the user requirements specifications should be addressed in the functional requirements.

- The user should be able to register and manage his appointments online at any time.
- Database has to store all the information efficiently without any information loss.
- The user shall be able to search for the doctors by specialty, name, working time and/or gender.
- The user can change his profile info at any time
- hospital can manage all appointments made with him on his account

6.TYPES OF TESTING

Functional Test

Functional tests provide systematic demonstrations that functions are tested as specified by the business and technical requirements, system documentation, and user manuals.
Functional testing is done based on the following things:
Valid Input : the identified classes of valid input should be accepted.
Invalid Input : the identified classes of invalid input should be rejected.
Functions : identified functions are to be exercised.
Output : identified classes of application outputs should be exercised.

The preparation of functional tests is focused on requirements, key functions, or special test cases. In addition to this, systematic coverage is done to identify Business process flows; data fields, predefined processes, and successive processes must be considered to do the 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 must be verified. The testing of all individual software units of the application is done after the completion of an individual unit before integration. It is a structural testing, that mostly relies on the knowledge of its construction and is invasive. Unit tests perform basic tests that are at component level and test a specific business process application, and system configuration. Unit tests ensure that each and every path of a business process performs accurately to the documented specifications and contains very clearly defined inputs and expected outputs

Figure 3: User login page

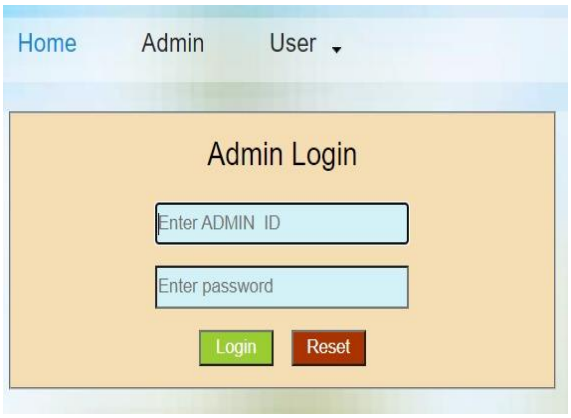

Figure 4: Admin login page

7. RESULTS

Figure 2: User registration page

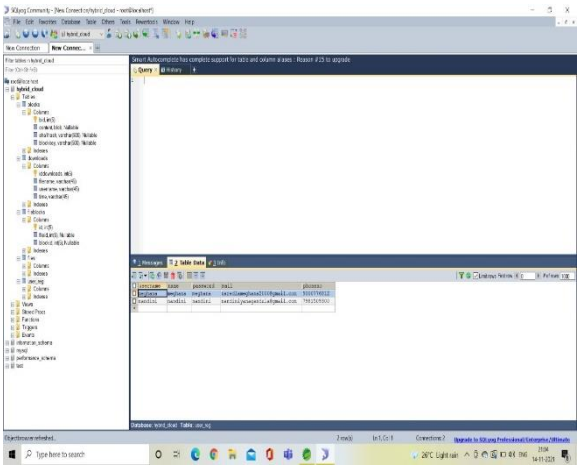

Figure 5: Registered users

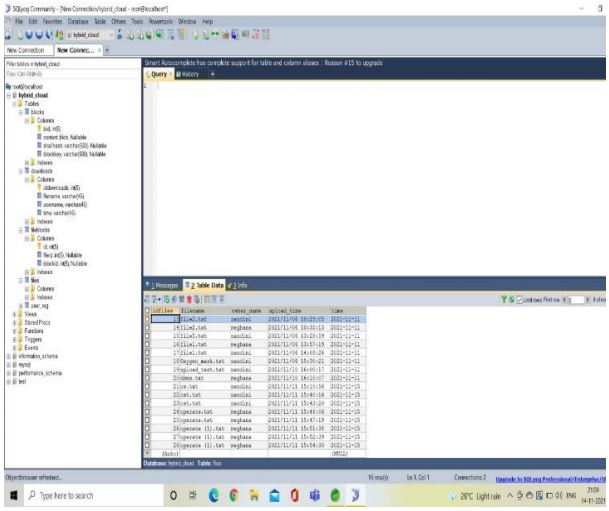

Figure 6: Uploaded files

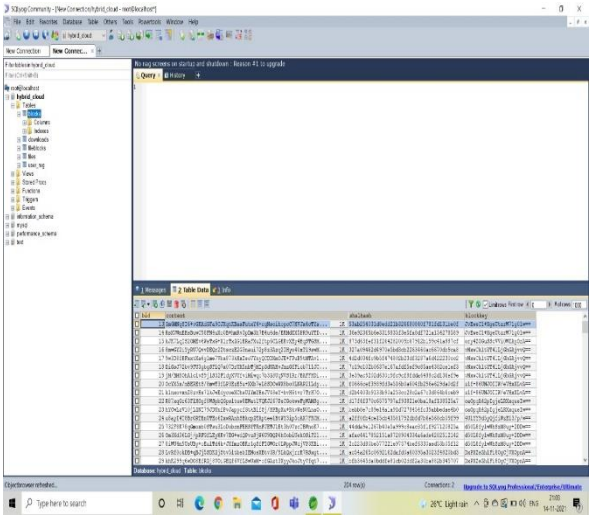

Figure 8: Generating hash values for blocks of data

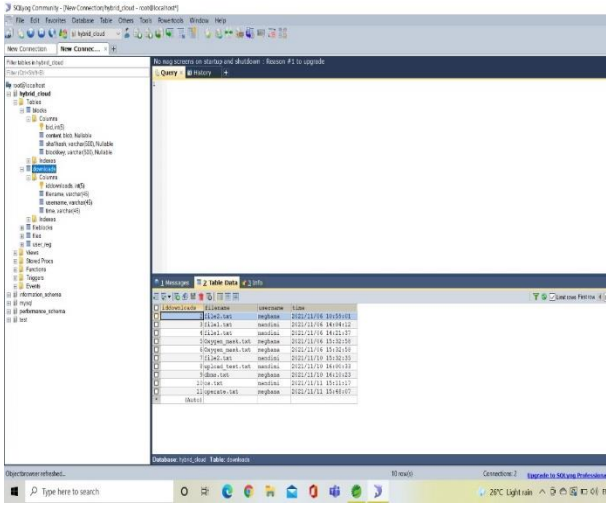

Figure 7: Downloaded files

8. CONCLUSION

- Data deduplication tremendously improves the efficiency of disk-based backup and decreases the quantity of deposited information.
- this deduplication scheme is simple and robust against the attacks while remaining efficient in terms of storage space and bandwidth.
- The hash key value is generated for the blocks of data through which data duplication is removed.
- Our method provides protection against attacks from malicious clients.

9. FUTURE SCOPE

- It reduces the requirements in cloud storage and also manages the volume of data that transfers through the network.

- It provides rapid results and improves data protection operations making them more efficient.
- It is cost effective as it requires fewer disks the storage cost is reduced.
- Eliminating the unwanted use of network bandwidth.

10. REFERENCES

- https://www.mysql.com/
- https://www.w3schools.com/
- http://tomcat.apache.org/
- https://www.tutorialspoint.com/webdevelopmenttutorials.htm
- https://www.javatpoint.com/servlet-api
- D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," in Proceedings of the 9th USENIX Conference on File and Storage Technologies (FAST), 2011.
- M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proceedings of the 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2013.