

Hybrid Cloud Data Protection Using Machine Learning Approach



D. Praveena, S. Thanga Ramya, V. P. Gladis Pushparathi, Pratap Bethi, and S. Poopandian

Abstract In today's digital world, information created by Internet of Things (IoT) devices has expanded drastically. This expansion is due to an increase in the number of IoT devices associated with the internet. Hybrid cloud computing provides enormous support to these emerging IoT devices in processing vast data. However, security is a challenging issue because of the integration of IoT and hybrid cloud. To achieve a sufficient level of hybrid cloud IoT security, a combination of Enhanced C4.5 machine learning algorithm and Dynamic Spatio Role-Based Access Control Algorithm is introduced. In this approach, the data users are classified using the Enhanced C4.5 algorithm and the user's level of cloud data access is restricted using the Dynamic Spatio Role-Based Access Control Algorithm. As a result, the major security issues pertaining to IoT cloud are addressed. The security framework also uses a deduplication algorithm for eliminating redundant data and significantly improving IoT data storage requirements.

Keywords Internet of Things (IoT) · Machine learning · Hybrid cloud networks · Security · Deduplication · Access control · Secured storage

D. Praveena (✉) · S. Thanga Ramya
Department of Information Technology, R.M.D Engineering College, Kavaraipettai 601206, India
e-mail: praveena.it@rmd.ac.in

S. Thanga Ramya
e-mail: str.it@rmd.ac.in

V. P. Gladis Pushparathi
Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai, India

P. Bethi
OPG Power Generation Pvt. Ltd., Chennai, India

S. Poopandian
Infosys Pvt., Ltd., Chennai, India

1 Introduction

The major concern for hybrid cloud computing is security challenges despite the advantages of on demand self-service, pay per use, vast network access and resource elasticity [1]. To provide security to the data, existing mechanism encrypts the data before the data is stored in cloud service provider. However, the same data can be duplicated by different users with different encrypted values at the cloud. In spite of the vast cloud storage space, duplication waste resources, power and increases the complexity of management of data. Thus deduplication has become necessary task for efficient cloud storage. Deduplication effectively save cost and space up to 90–95%, thereby reducing the burden of the cloud users.

Deduplication over the encrypted data is provided by convergent encryption. In convergent encryption (CE), the encryption key is obtained by hashing the input file. Then the encryption key is used to encrypt data and gets cipher text for cloud server storage. Similar cipher text is yielded for the same file hash value; thus deduplication is performed. However, convergent encryption is subject to offline brute-force dictionary attacks [2]. DupLess technique avoids brute force attack and ensures high security as long as the key server which aids in data encryption is secured from the attacks.

Convergent encryption security threats can also be solved by Cloud Dedup [3] but still it cannot handle the issues of data deletion. An information holder that expels the information from the cloud can even now get to similar information since it despite everything knows the information encryption key if the information isn't totally expelled from the cloud. The existing solutions the reduces the storage overhead to greater extent but still data users find in difficulty access control in data storage in a convenient way.

2 Problem Statement

Cloud security is still a hard issue in recent days particularly for the hybrid cloud because of the mixture of private and public cloud. Many existing security mechanisms available are heavy time consuming, privacy issues and lacks flexible access control [4]. To solve the major hybrid cloud issues, a new approach is proposed which uses a machine learning technique for efficient data storage and retrieval. This approach uses newly proposed deduplication algorithm with Dynamic Spatial Role Based Access Control and enhanced C4.5 machine learning algorithm [5] for restricting unauthorized user accessing the cloud without redundancy. The user can download the requested data in a flexible manner, once the user access rights are verified.

3 De-duplication

Deduplication, also called single instance technique, is a compression technique for expelling duplicate data and storing the original data copy with the goal that it will save the storage space to protect sensitive data [6]. The information security and access to specific data are particularly vital in recent times. Therefore, the features in deduplication have been generally used in the cloud storage system [2]. The management of the consistently expanding volume of data is critically tricky for cloud storage system; the data management makes conceivable through deduplication concept. In the organizations, the storage systems of several part contain multiple copies of same data [7]. The different users save the same file might at various places, which means multiple copies are getting created which has the similar data. Deduplication dispenses these multiple copies which get created by storing a data copy and a pointer pointing to the other copies of the actual file. Deduplication technique is utilized in organizations for disaster recovery and backup applications; this strategy also used to make free space in the storage systems [6].

Deduplication lessens the needs of storage by up to 90–95% for the application of reinforcement, 68% in standard file framework and therefore we generate more space for our file [8]. Essential problems in data deduplication are to give privacy and security to shield the data from insider or outsider attacker. Figure 1 shows the deduplication process before accumulating data into the memory.

The two major types of deduplication, depending on granularity are block-level deduplication and deduplication based on file-level. The entire file is considered for file-level deduplication, thereby a small change or update creates a new file which is different from older version of it, resulting in reduction of duplication ratio. But data blocks are considered for deduplication in block-level deduplication.

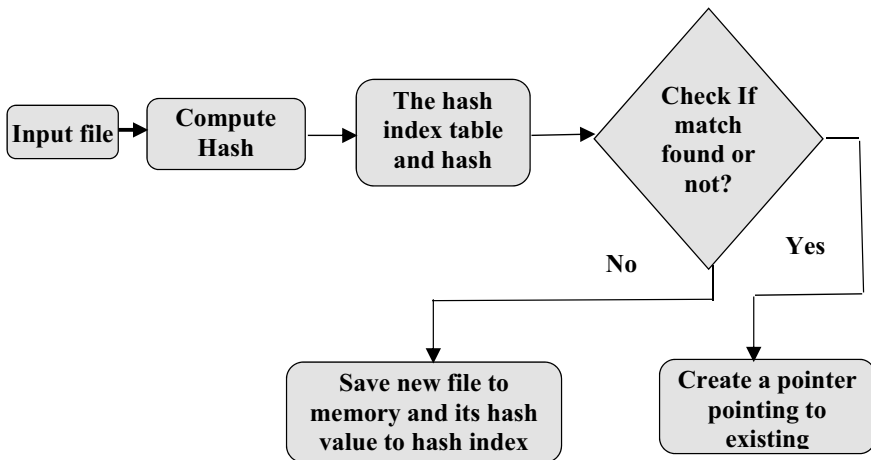


Fig. 1 Data deduplication flowchart

Location of deduplication can also be considered for dividing the deduplication, i.e., as source-side deduplication and as client-side deduplication. Bandwidth is guaranty saved in client-side deduplication guarantees as only file's hash value is transferred to the server if the duplicate exists [9].

Deduplication is most utilized in different applications like primary storage, backup, metadata management and so on for storage optimization [10]. Data deduplication has turned into an essential part of controlling repositories of external data to the cloud data centres. In any case, centralized data centres face issues of data loss and accessibility if something goes flawed as deduplication keeps up only a unique copy of the content.

When a user needs to download a document from the cloud, he can only login into the system and select file for download that is in encrypted file form, so user decrypts this file utilizing convergent key. To make information administration flexible in distributed computing, deduplication has been an understood system and has become much thoughtful at recent times [8]. Information deduplication is a technique applied on certain information and pressure procedure for removing copy duplicates of rehashing information. The procedure is employed to improve stockpiling us-age and can similarly be associated with network information transferring to decrease the quantity of bytes that must be sent. Deduplication eliminates bounty information, by putting away one duplicate of physical information and diverting other recreate information to that physical duplicate without keeping different duplicates of same information [11].

4 Cloud Security Protection Framework with Machine Learning Modules

The overall framework of the proposed method is shown in Fig. 2. The security system has four diverse subcomponents to be specific access control stage, de-duplication stage, area guide chief. The information assortment module has gathered the cloud client information and the association's reports and put away in the hybrid cloud storage. The duplicate data is expelled, and the scrambled de-duplication module encodes the accessible information that are as of now put away in the hybrid cloud.

The hybrid cloud data recovery is the responsibility of encrypted reduplication, that are as of now accessible in the cloud storage and evacuating the duplication data and furthermore in encrypting the recovered information.

The UI block is responsible for storing enciphered data in the hybrid cloud storage without replication. From the rundown of security dangers, redundancy is one of the noteworthy dangers. The cloud user stores the organization data in the hybrid cloud database via the user interface module with approved authorization control of access control. The cloud users are restricted in accessing the cloud storage by efficient access control module. The module permits only the authorized user of the organization with approved credentials.

Access control algorithm and location manager are the two components of the access restriction control system in the security framework. The position administrator shows the cloud clients area wherefrom they are getting to the information by different cloud clients. Here, the UI module acts as an interface and is in charge for communicating with all the modules of security framework, for example, cloud storage, encrypted de-duplication module, and components of access control management.

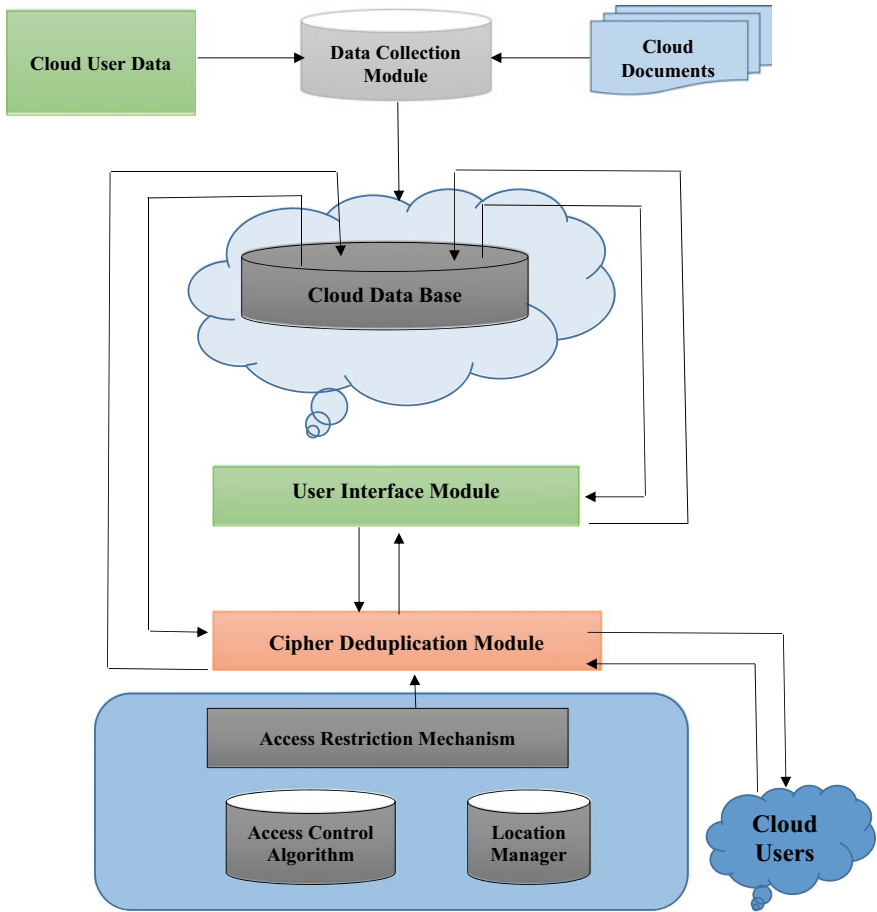


Fig. 2 Hybrid cloud data protection with a machine learning application

4.1 *Cloud Client Classification Using Enhanced C4.5 Algorithm*

Maintaining a security management of person authorizations is an essential element of system. In a multi-area or a dynamic environment, those essential constructing modules of get entry access to manage aren't so well hooked up. RBAC nevertheless lacks the capacity to address evolving context element and managing any unexpected situations [12]. For example, whilst a brand new person requests a get right of entry to and sources they want may also vary, conventional get right of entry to manage mechanisms lacks capacity to address those form of requests effectively. In this situation, it has to now no longer be vital to absolutely extrude the hooked up local get right of entry to manage mechanisms for a new client. Instead, to aid dynamic environments, it has to be viable to automate the era and enforcement of get right of entry to manage guidelines to satisfy the new user needs, even as the existing mechanism continue to be intact.

Normally, when another client is added to a current framework, manual intercession of security director gets important to choose whether a client should access a mentioned asset or not. In a role based access control security condition, if the user's access rights are still unknown or the mentioned access' point of view are not significant, it gets hard for the security overseer to decide a reasonable job for another client through which a client can get to benefits that are required to play out a specific activity [13]. From security point of view, it is critical to guarantee that conceded access to another client is steady with the hierarchical access control strategy. Be that as it may, a security overseer can commit an error in allotting proper job to another client. In this manner, when a RBAC is set up, deciding a reasonable job for another client is a basic yet testing task. As shown in Fig. 3, the UI module has a current AI technique known as Improved C4.5.

The UI module has a current AI technique known as Improved C4.5 scheme [14] to characterize the cloud users who request information from the cloud database. The model grouping of a user for any role is developed by utilizing the data about permission of existing clients who have that role, and furthermore the qualifications of existing clients who don't have that role. As it were, labelled training information (supervised learning) for roles are first utilized for model structure process and later those models are utilized to characterize a client for a possible role.

A few classification techniques have been proposed in the writing of data mining and AI. Decision tree is ordinarily utilized as a choice help apparatus that utilizes a tree-like chart or model of choices and conceivable result of those choices for the choice investigation reason. Improved C4.5 calculations depends on isolate and overcome technique where issue of taking in arrangement structure from known arrangement of perception prompts a methodical development of a decision tree.

The existing system involves three steps to add a new user to the overall process:

1. For each role, build a classification model.
2. Selecting roles having requested permission.
3. For each role external user's attributes are classified.

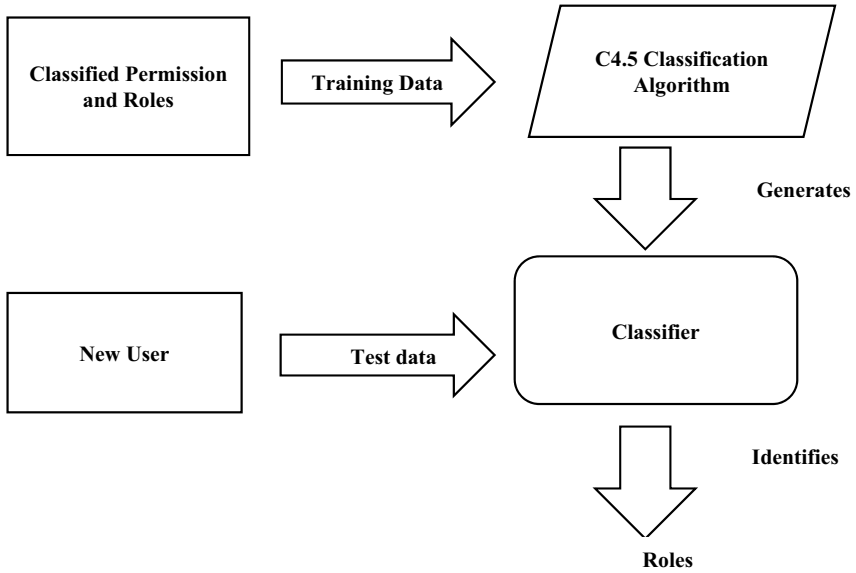


Fig. 3 User role classification process using C4.5 algorithm

C 4.5 Algorithm for assigning role to a new user

Input: PA: Permission role assignment

Output: Assignment of role to new user_role

Step 1: new user_role → empty

Step 2: for each role in PA do

- Step 2.1: If the role has permission on a file then
- Step 2.2: new user_role → role

Step 5: End

4.2 De-duplication Processing Algorithm

The novel de-duplication algorithm removes the replicated contents that are existing in the hybrid cloud. This enables access in matching and eliminating the similar

files that are available in the hybrid cloud storage. The cryptographic features in the deduplication processing algorithm are utilized for the encryption of the data.

Encrypted Data Upload: Data owner encrypts the file to be uploaded using randomly selected symmetric key using AES-256 algorithm and stores only the encrypted file in the hybrid cloud storage. Further for each file the CSP computes the hash token with SHA-512 for the file and stores hash token along with encrypted symmetric key as record in cloud storage [15]. Later the data owner can share the encrypted symmetric key for the valid data user using RSA public key cryptography.

Data De-duplication: When the same files are to be stored in the cloud storage, the data deduplication occurs. CSP compares the hash token for the new file with the records in the cloud data [16]. If the file already exists, the CSP checks the access rights of the data holder and if valid permits the symmetric key to decrypt the encrypted data.

Data Updation: The data owner updates encryption key to all the data user, whenever the new update encryption is updated.

Data Deletion: When CSP receives the data deletion request from user A, the CSP checks the eligibility, then deletes the record of A and its access rights to the file. Further it checks that the deduplication records are empty, if it is empty, it removes the encrypted data and all its records.

4.3 Dynamic Spatial RBAC Algorithm

Dynamic Spatial RBAC Algorithm provides to restricted and secure access to the cloud users for accessing the hybrid cloud files. DSRBACA is capable of assigning a set of roles described by set of permission to different cloud users dynamically under various constraints. Permissions are approval to execute set of operation on one or more data and depends on the role and owner location. This role is periodically changed the location of the cloud user is identified and monitored dynamically for fixing their rights to access their data organization.

The user permissions are validated with the access rights and in addition, the spatio constraints with the status level of the user are used for validating spatio action status access control policies [17]. Status level is a named position of a user with other users in a specific domain. For example, a bank manager and clerk has two distinct status levels in bank domain where the manager status level is higher than clerk in accessing files. In DSRBACA, a user's status level is determined from the ascribed status and action status at a particular location. The assignment of user status level and access privileges changes dynamically with respect to the change in the location.

Algorithm for Deduplication Process***First Phase: Data Upload***

- Step 1: Calculate the hash token for the user input file using SHA-512 algorithm
- Step 2: Apply symmetric encryption algorithm AES-256 to encrypt the file with key k
- Step 3: Apply RSA algorithm to encrypt the symmetric key K used in above step
- Step 4: Store the encrypted input file into the database along with hash token and encrypted key.
- Step 5: End

Second Phase: Deduplication

- Step 1: Calculate the hash token for the files from the database
- Step 2: Check whether hash token already exist in the database
- Step 3: It checks whether a different user is storing the same data, if yes, the CSP performs deduplication,
- Step 4: If no, hash token of the encrypted content along with the key will be stored in the cloud database.
- Step 5: End

The steps involved in Dynamic Spatial RBAC Algorithm are as follows:

Dynamic Spatial RBAC Algorithm

```

1: Read → cloud users requirement
2: user requirements are initialized in queue
3: For accessing the data, Read → cloud users requirement
4: While (! isEmpty(Queue))
do
5: From the queue find the next user
6: Verify → cloud user query by using the database access privileges
and the user level
7: Check whether (user level > 5), then
8: Verify → access rights → use time and location to assign the suitable role
9: Update user level based on the time and location
10: If (check the time and spatial constraints to validate cloud user rights)
11: Check if (valid cloud user), then
    Find the address of the data stored and its position in the hash table
    Input → respective position with content and key
    Decrypt → the stored encrypted data with RSA
    Show → original message
Else
    Show → ERROR.
12: Repeat → until 4 to 11 until the no user in queue.
13: End

```

Dynamic Spatial RBAC Algorithm utilizes RSA data encryption and the cloud datasets for time and spatial requirements is assessed by the implementation for spatio access control mechanism.

5 Experimental Results

An OpenStack cloud is used to implement the security system framework and quantities of trials are assessed. In implementation, the trials were led utilizing a 2.4 GHz Intel I5 processor running at, 7200 RPM Western Digital, 4 GB RAM, and 8 MB cradle 350 GB Serial ATA drive. It includes, information examining process was completed to assess the exhibition of the framework. In the results five different

Table 1 Analysis of users request denial

Sl. No.	Count of tried user request	Count of requests denied by	
		DDPA + DSRBACA	RBAC
1	100	6	5
2	200	10	9
3	300	15	13
4	400	19	17
5	500	20	22

values analysed to test clients accessing the cloud storage and access denial in the proposed framework. For this reason, we have attempted an alternate number of client demands, for example, 100, 200, 300, 400 and 500 of every five analyses. Table 1 shows the client demands denied examination. Additionally, the solicitations incorporated the certifiable and the malignant client demand with an extent 19:1.

In Table 1, the results show that the Dynamic Spatial RBAC Algorithm and the ordinary RBAC model were compared in confining the cloud user quantity, furthermore gives over 90% location and anticipation precision. This is on the grounds that the utilization of worldly and spatial requirements while settling on a choice. Successful and made sure about proposed deduplication is utilized.

The performance of clients’ request wherein the quantity of approved clients in the DSRBAC algorithm is shown in Fig. 4.

In Fig. 4, the DSRBACA authorization access is less than the existing RBAC algorithm. In addition, cloud users less than 5% in the current framework were denied access, and subsequently the RSA encryption algorithm upgrades the security. This

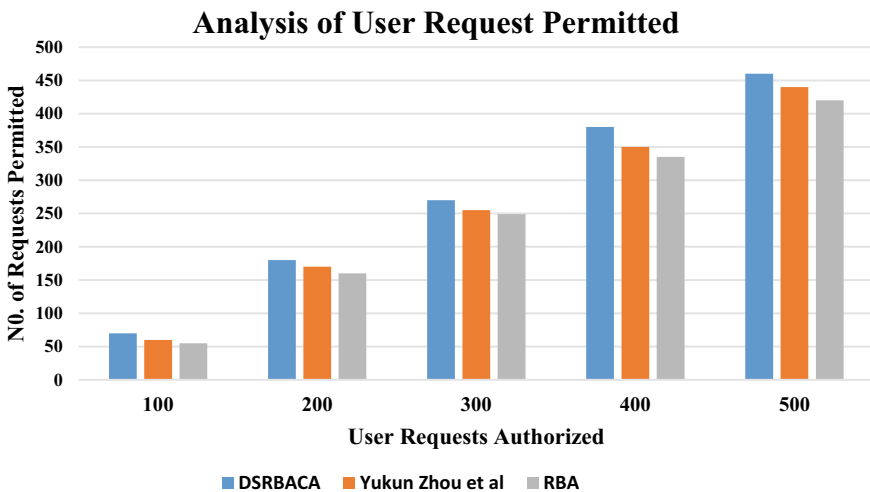


Fig. 4 Analysis of permission on user request by RBAC and DSRBACA

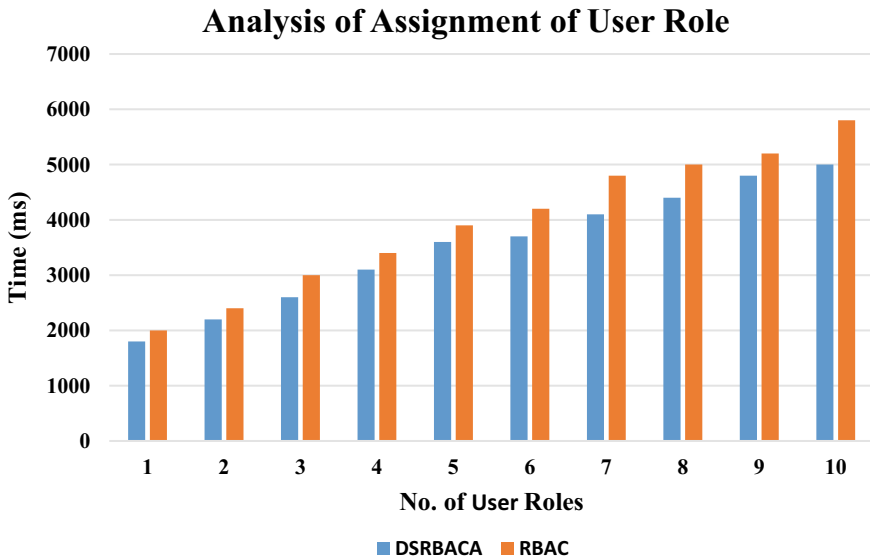


Fig. 5 Analysis of assignment of user role

is a direct result of the utilization of existence requirements viably for distinguishing the irregular clients.

The analyses directed needs 959 ms to make 10 roles, 9 edges are incorporated in 844 and 711 ms to concede 10 permissions for every one of the t roles, and 3,384 ms to dole out 50 clients to every job. The time involved for every one of the basic activities is in the range 68–120 ms. The time taken to dole out ten clients with ten roles is shown in Fig. 5.

DSRBACA and RBAC were compared in terms of the time taken to process the constraints in Fig. 5, graph shows that DSRBACA takes less time than RBAC.

Figure 6 examines the approval time examination of the proposed DSRBACA, RBAC and renouncement algorithm [18]. Here, we have led five examinations with an alternate number of documents, for example, 100, 200, 300, 400 and 500 MB.

Figure 6 depicts the enhanced security performance of the DSRBACA, standard RBAC and Zhou et al. [19].

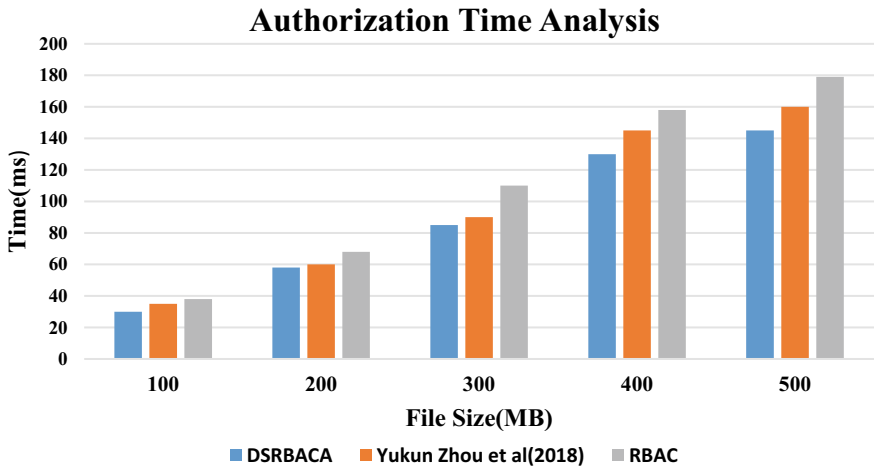


Fig. 6 Authorization time analysis

6 Summary

Despite the fact that all the essential activities are insufficient for cloud storage and recovery. A large quantity of data in cloud storage application is distributed in numerous locations. These drawbacks have been overwhelmed by the recently proposed Dynamic Spatial RBAC Algorithm is dynamic dependent on time and location. This mix of location and time are useful for additional improvement of the cloud data security. DSRBACA has constrained the access of client’s information or records that in the hybrid cloud storage system. In correlation with the RBAC model in limiting the quantity of cloud clients, the proposed DSRBACA performs well and furthermore gives over 90% location and avoidance precision.

Moreover, approved cloud users less than 5% were denied access in the current framework and consequently the RSA encryption algorithm upgrades the security with better confidentiality of data. Generally speaking, the security dangers in the hybrid cloud systems were improved utilizing cryptography algorithms.

References

1. Praveena, D., Rangarajan, P.: Analysis of trend, service and deployment models in cloud computing with focus on hybrid cloud and its implementation. *Res. J. Appl. Sci.* **9**(4), 181–186 (2014)
2. Bellare, M., Keelveedhi, S., Ristenpart, T.: Message locked encryption and secure deduplication. In: *Proceedings in 32nd Annual International Conference on Theory Application Cryptographic Technology*, pp. 296–312 (2013)
3. Yuan, H., Chen, X., Jiang, T., Zhang, X., Yan, Z., Xiang, Y.: DedupDUM: secure and scalable data deduplication with dynamic user management. *Inf. Sci.* **456**, 159–173 (2018)

4. Praveena, D., Rangarajan, P.: Improving the security in web application cloud service Karatsuba Montgomery multiplier ECC algorithm. *J. Web Eng.* **17**(6), 3078–3094 (2018)
5. Thanga Ramya, S., Arunagiri, B., Rangarajan, P.: Novel effective X-path particle swarm optimization based deprived video data retrieval for smart city. *Clust. Comput.* 13085–13094 (2019)
6. Shete, A., Patil, B.M.: Deduplication in hybrid cloud with secure data. *Int. J. Comput. Appl.* **148**(8), 32–36 (2016)
7. Raji, L., Thanga Ramya, S., Thilagavathy, A.: Reducing cloud storage space consumption using PRCR. *Int. J. Innov. Technol. Explor. Eng. (TM)* **46**, 287–289 (2019)
8. Zhou, L., Varadharajan, V., Hitchens, M.: Achieving secure role-based access control on encrypted data in cloud storage. *IEEE Trans. Inf. Forensics Secur.* **8**(12), 1947–1960 (2013)
9. Akhila, K., Ganesh, A., Sunitha, C.: A study on deduplication techniques over encrypted data. *Procedia Comput. Sci.* **87**, 38–43 (2016)
10. Meyer, D.T., Bolosky, W.J.: A study of practical deduplication. *ACM Transactions on Storage (TOS)* **7**(4), 1–20 (2012)
11. Gapat, P., Khillare, S., Khiste, A., Pise, R., Gapat, P.: Securing data deduplication on hybrid cloud using asymmetric key algorithm. *J. Innov. Res. Sci. Technol.* **2**, 759–763 (2017)
12. Praveena, D., Rangarajan, P.: A machine learning application for reducing the security risks in hybrid cloud networks. *Multimedia Tools Appl.* **79**(7), 5161–5173 (2020)
13. Qiu, M., Gai, K., Singham, B.T., Tao, L., Zhao, H.: Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry. *Future Gener. Comput. Syst.* **80**, 421–429 (2018)
14. Prema Rajeswari, L., Arputharaj, K.: An active rule approach for network intrusion detection with enhanced C4.5 Algorithm. *Int. J. Commun. Netw. Syst. Sci.* **1**(4), 314–321 (2008)
15. Yan, Z., Wang, M., Li, Y., Vasilakos, A.V.: Encrypted data management with deduplication in cloud computing. *J. Cloud Comput.* **3**(2), 28–35 (2016)
16. Zhenhua, L., Yaqian, K., Chen, L., Yaqing, F.: Hybrid cloud approach for block-level deduplication and searchable encryption in a large universe. *J. China Univ. Posts Telecommun.* **24**(5), 23–34 (2017)
17. Helmi, A.M., Farhan, M.S., Nasr, M.M.: A framework for integrating geospatial information systems and hybrid cloud computing. *Comput. Electr. Eng.* **67**, 145–158 (2018)
18. Li, J., Li, Y.K., Chen, X., Lee, P.P.C., Lou, W.: A hybrid cloud approach for secure authorized deduplication. *IEEE Trans. Parallel Distrib. Syst.* **26**(5), 1206–1216 (2014)
19. Zhou, Y., Dan Feng, Yu., Hua, W.X., Min, F., Huang, F., Zhang, Y.: A similarity-aware encrypted deduplication scheme with flexible access control in the cloud. *Future Gener. Comput. Syst.* **84**, 177–189 (2018)



Dr. D. Praveena B.E., M.E., Ph.D., is an Assistant Professor in Department of Information Technology, RMD Engineering College, since August 2006. She has completed his Ph.D. (Hybrid Cloud Security) at College of Engineering, Anna University in the year 2019. She has been in the teaching profession for the past 15 years and has handled UG programmes and published 7 papers in various International Journals and Conferences.



Dr. S. Thanga Ramya B.E., M.S. (by Res), Ph.D., is an Associate Professor in the Department of Information Technology, since June 2008. She obtained her B.E. (CSE) from Dr. Sivanthi Aditanar College of Engineering and M.S. by Research (ICE) from Anna University, Chennai. She has obtained her Ph.D. in Information and Communication Engineering from Anna University, Chennai, in 2017. She has been in the teaching profession for the past 18.5 years and has handled both UG and PG programs. Her areas of interest include programming languages, database management and data mining. She has published 13 papers in various International Journals and Conferences. She has also published four text books for the benefit of B.E./B.Tech. students of Anna University. She has got Oracle java international certification, IBM RAD, TIVOLI and DB2 certification and obtained her PRP certification from Wipro. She secured bronze medal from Infosys for the campus connect program. She is the life member of ISTE.



Dr. V. P. Gladis Pushparathi B.Tech., M.Tech., Ph.D., is an Associate Professor in Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, since Dec 2016. She obtained her B.Tech. from Cape Institute of Technology, Tirunelveli and M.Tech. from M.S. University Tirunelveli. She has obtained her Ph.D. in Information and Communication Engineering from Anna University, Chennai, in 2016. She has been in the teaching profession for the past 14 years and has handled both UG and PG programs. She has completed major and minor funded projects. She has published 21 papers in various International Journals and 78 papers in international Conferences. She has published more than 15 Indian patents. She has also published three text books for the benefit of B.E./B.Tech. students of Anna University. She has got best faculty award from Neghru Group of Institutions. She is the life member of ISTE, IACSIT, IAENG and ISRD. She is the reviewer for Journal of computer and Electrical Engineering/Elsevier and Journal of 3D Research. Her field of Interest includes Digital Image Processing, Soft Computing, and Data mining.



Mr. Pratap Bethi working as General Manager (SAP & IT) in OPG Power Generation Pvt. Ltd. He has around 17 years of experience in SAP and other verticals. He completed Bachelor's degree in Technology (B.Tech.) with the specialization in Mechanical Engineering from J.N.T.U. in April' 2003. His area of interest is ERP/Integration of third party systems with SAP and Cloud Technologies.



Mr. S. Poopandian is a principal consultant in Infosys Pvt., Ltd., He has 16 years of experience in SAP related project (Building Global Template, Implementation of Product costing, COPA International Rollout Programs, AMS Lead), SAP Implementations, Rollouts and Upgrade. He has working experience in multicultural environments (Asia, US, Europe, Africa, Asia). He is SAP Management Accounting Certified Solution Consultant ECC 6.0. He Worked on S4 Hana on POC for Various clients. He is having expert knowledge in SAP RAR solution for IFRS 15 Compliance. He has done couple of India CIN Implementation and 3 India GST Implementation for Global Automobile Majors, Wind power energy producers and Car Chassis Manufacturers. His area of interest mainly focuses on cloud computing.