SECURE BLOCK-LEVEL DATA DEDUPLICATION APPROACH FOR CLOUD DATA CENTER

Mr.K Somanatha Rao[1], Ms. Neha Hasan[2], Mr. Khaja Pasha Shaik[3]

ksrao@lords.ac.in

[1,2,3]Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

Abstract—

The on-going growth in information and technology sector has increased storage requirement in cloud data centres with unprecedented pace. Global storage reached 2.8 trillion GB as per EMC Digital Universe study 2012 and will reach 5247GB per user by 2020. Data redundancy is one of the root factors in storage scarcity because clients upload data without knowing the content available on the server. Ponemon Institute detected 18% redundant data in “National Survey on Data Centers Outages” . To resolve this issue, the concept of data deduplication is used, where each file has a unique hash identifier that changes with the content of the file. If a client tries to save duplicate of an existing file, he/she receives a pointer for retrieving the existing file. In this way, data DE duplication helps in storage reduction and Identifying redundant copies of the same files stored at data centres. Therefore, many popular cloud storage vendors like Amazon, Google Drop box, IBM Cloud, Microsoft Azure, Spider Oak, Walla and Mazy adopted data DE duplication. In this study, we have made a comparison of commonly used File-level DE duplication with our proposed Block-level DE duplication for cloud data centres. We implemented the two DE duplication approaches on a local dataset and demonstrated that the proposed Block-level DE duplication approach shows 5% better results as compared to the File-level DE duplication approach. Furthermore, we expect that the performance will further be improved by considering a large dataset with more users working in similar domain.

I. INTRODUCTION

Cloud services provide almost unlimited storage capacity, allowing customers to use as much space as they need. Vendors continuously seek methods to reduce duplicate data (many copies) and enhance storage efficiency. In order to reduce the presence of repetitive data, which refers to the removal of numerous duplicates, we use the deduplication approach. Cross-user deduplication is the approach that is most often used. The fundamental concept of deduplication is to save redundant data just once. Consequently, if a user intends to upload data that is already stored in the cloud, the cloud provider will display a message indicating that deduplication is not permitted. Deduplication may significantly decrease storage requirements, achieving a reduction of up to 90-95% for backup applications [11] and up to 68% in conventional file systems [23].Users want data security, secrecy, cost-effectiveness, and adaptability, all of which are guaranteed with encryption. Regrettably, deduplication and encryption are inherently incompatible technologies. Deduplication aims to identify and save duplicate data only once, while encryption renders two identical sets of data indistinguishable after being encrypted. Consequently, if customers encrypt data using a standardized method, the cloud storage provider is unable to do deduplication since identical data would seem distinct after encryption. However, if customers do not encrypt their data, cloud storage companies cannot provide confidentiality and protection against attackers. Convergent encryption. Convergent encryption is a suggested solution that aims to address the conflicting needs by generating the encryption key from the hash of the data. Convergent encryption seems to be a promising solution for achieving both secrecy and deduplication simultaneously. However, it is unfortunate that this method is plagued by many well-known vulnerabilities [15], [24]. Our primary emphasis is on deduplication and cloud storage. Cloud computing revolutionizes the delivery of Information Technology services by efficiently organizing and distributing resources such as storage and processing power to users according to their specific needs.

Furthermore, the relinquishment of control over one's personal data results in elevated data security hazards, particularly in terms of data privacy breaches. The privacy concern has escalated significantly as a result of the fast advancement of data mining and other analytical technologies. Therefore, it is advisable to only delegate encrypted data to the cloud in order to guarantee data security and protect user privacy. However, it is possible for either the same or separate users to upload replicated data in an encrypted format to a Cloud Service Provider (CSP), particularly in situations when data is shared among several users. While cloud storage capacity is extensive, data duplication significantly squanders network resources, uses substantial amounts of energy, and adds complexity to data management. The proliferation of various services will intensify the need to implement effective resource management techniques. The primary concern is the effective management of encrypted data storage while using deduplication. Nevertheless, existing industrial deduplication methods are incapable of processing encrypted data. Current deduplication systems exhibit security vulnerabilities, such as susceptibility to brute-force attacks.

Simultaneously, they lack the ability to effectively accommodate both data access control and revocation in a flexible manner. The majority of current solutions lack the capacity to guarantee dependability, security, and privacy. Implementing deduplication management for data holders is challenging in reality owing to several factors. Initially, they might potentially result in storage delays due to the fact that data holders may not always be online or accessible for such a management. Furthermore, the procedure of including data holders in the deduplication process may become too intricate in terms of communications and calculations. Furthermore, throughout the process of identifying duplicated data, there is a potential invasion of privacy. Furthermore, in some scenarios when a data holder lacks knowledge about other data holders owing to data super distribution, it may be unaware of how to provide data access rights or assign deduplication keys to a user. Consequently, CSP is unable to collaborate with data holders in several scenarios for the purpose of data storage deduplication. The findings demonstrate the scheme's exceptional efficiency and efficacy, particularly for the actual implementation of data deduplication in cloud storage.

II RELEATED WORK

Recent advances in the stock markets have caused significant effects on finance which can be more complicated to predict the indexes. Nowadays, most of people are directly or indirectly related to this subject and the more technology is developing the more they need to know and predict the indexes and this makes them be interested in index prediction. However, due to the quick changes in stock price, the prediction of stock price becomes a challenging task. Moreover, effects of the cryptocurrencies have increased this complexity [1]. These factors cause traders to go through using intelligent systems rather than using fundamental analysis to predict the price. Accordingly, traders can sell the index before value decline or buy before the price rises and this causes the trader to have much more profit. Also, it seems unbelievable for traders to replace their experience and professionalism with intelligent systems, but due to the remarkable amount of data and technological advancements of intelligent systems, algorithms, pattern recognition and Artificial Intelligence, it seems appropriate to use and even, combine them with the experience and professionalism. Since the significance of accurate information, Neural Networks (NN) have become one of the successful and efficient algorithms and models that are being used for modeling stock market behavior [2].

Artificial Neural Network (ANN) is a popular method which also incorporate technical analysis for making predictions in financial markets. One of the most practical methods in this area is Long Short Term Memory (LSTM) [3]. Pattern recognition is another method to predict the repeated patterns of a stock index over a period of time in the future [4]. In stock trading, it is very decorous that a model like NN provides a prediction nearly to the real price. Predicting the stock market is one of the processes that requires experience and reacquaint to have an accurate prediction. However, this process is qualitative and cannot be a complete prediction. AI and NN convert it to a quantity that means it can be used with mathematical approaches and results in a scalar number form that gives an amount with high accuracy and small Root Mean Squared Error (RMSE) values for future that is more reliable than qualitative predictions[5, 6]. Although this approach is useful, all users cannot apply it due to it is not implemented on an OS program. One of the Neural Networks is Real-Time Recurrent Learning (RTRL) network that is practical and able to store the information for later use to enhance the efficiency and a better way for modeling [7]. In [8] it is proved that the algorithms based on RNNs can be useful in financial market prediction.

The program which is presented in this paper can be flexible and dynamic that means user can change the mark in fraction of a minute. When it is incorporated with Python programming AI can be more advantageous than previous methods, and a wide range of people can use it. The NN is means of performing machine learning, in which a computer learns to perform some tasks by analysing training examples, and that is the base case of prediction, and the approach to quantitatively predict the stock market. This simple feature is used in this paper to make complicated approaches for accurate prediction close to the real value. The organization of the paper is as follows: In section 2, we provide preliminaries and tools. The proposed framework is presented in section 3. The results are presented in section 4. Finally, section 5 draws a conclusion and further improvements.

present incredible difficulties to successful investigation of feeling dispersion. In this approach a visual frameworks called Opinion Flow that enable experts to recognize feeling engendering designs and gather experiences. Enlivened by the data dispersion model and the hypothesis of specific presentation and build up a sentiment dissemination model to estimated feeling proliferation among Twitter clients. [6]

Cloud Architecture

Deduplication may be executed on many cloud architectures, including single cloud, multicloud, and hybrid cloud. Convergent keys and Proof of Ownership (PoW) technologies may be used in a single cloud architecture to safeguard data against loss and breaches. This strategy is widely used by the majority of commercial CSPs. In order to mitigate the risk of a single point of failure in a single cloud architecture, a multicloud architecture is used. This involves dividing the file into numerous shares and distributing them over different cloud servers. In order to prevent the need for disaster recovery in a multicloud environment, it is possible to implement efficient scheduling algorithms that provide both data reliability and quick recovery time.

Authorized data deduplication is carried out in hybrid cloud architecture [12]. This approach involves storing outsourced data on a public cloud while managing all data activities in a private cloud. The user has the ability to do a duplication check, provided that the user has the necessary rights. Secure deduplication may be achieved by encrypting the user file using distinct privilege keys.

SYSTEM ANALYSIS:

EXISTING SYSTEM:

Data redundancy is one of the root factors in storage scarcity because clients uploads data regardless of checking the same file already exists or not Cloud Service Providers adopting many simultaneous strategies including data DE duplication to overcome the possible scarcity of data storage. Data DE duplication has advocated a promising and effective role to save the digital storage space by removing the various copies of a data file available on data centers.

By maintaining just one duplicate of each file submitted, cloud storage service providers like Dropbox, Google Drive, Mozy, and others use deduplication to save space. Nonetheless, storage savings by deduplication are completely lost if customers encrypt their data in the traditional manner. This is due to the fact that distinct encryption keys are used to preserve the encrypted data as distinct contents. Current commercial technologies for deduplicating encrypted data are ineffective. One effective deduplication method, DeDu [17], for instance, is unable to process encrypted data. A client storing data at an untrusted server may use the current system's proven data possession (PDP) paradigm to confirm that the server indeed holds the original data without having to retrieve it. By randomly selecting blocks from the server, the approach produces probabilistic proofs of ownership while significantly lowering input/output (I/O) costs. The customer consistently keeps track of information in order to validate the evidence. The challenge/response protocol reduces network transmission by sending out a small, consistent quantity of data. Thus, big data sets in widely dispersed storage systems are supported by the PDP paradigm for remote data validation. We provide two provably-secure PDP schemes that outperform existing solutions, even when weighed against schemes that yield less robust guarantees.

DISADVANTAGES OF EXISTING SYSTEM:

- Time complexity
- Data sharing inflexibility due data privacy leakage.

PROPOSED SYSTEM: Deduplication technique can be mainly categorized into two types; File-level deduplication and Block-level deduplication. In File-level deduplication, the system generates the hash value of the complete file and then checks for the identical hash value in the hash-table. While Block-level deduplication approach, it divides the file in further segment called “chunks” and generates the hash value of each block to checks either its identical exist in the hash-table or not, Block-level deduplication is more efficient in the term of storage saving. Location wise where exactly deduplication takes place, it can be divided into two categories, according to the location, which are server-side deduplication and client-side deduplication.

In server side deduplication, client is unaware of the process and the deduplication occurs once file gets uploaded to the system, whereas in the client-side deduplication, client or uploaded is fully involved in the process of deduplication, initially client sends hash value of the desired uploading file greatly reduce the storage space and communication bandwidth. Therefore, many popular cloud storage vendors like Amazon, Google Drop box, IBM Cloud, Microsoft Azure, Spider Oak, Waula and Mozy adopted data deduplication technology. The commonly practiced procedure of data deduplication technology, it identifies the same data files or block through generating their cryptographic-hash string by using some hash function i.e., SHA-1, SHA-256 etc. Hash function generates the same hash string for the files having the same contents.

PROPOSED METHODOLOGY

In our proposed work Greedy and Dynamic Blocking Algorithms suggests tweets by coordinating clients with different clients having comparable interests. It gathers client input as evaluations gave by client to explicit tweets and discovers coordinate in rating practices among clients to discover gathering of clients having comparative inclinations. One of the principle highlights on the landing page of Twitter shows a rundown of top terms purported moving themes consistently. These terms mirror the points that are being talked about most at the exact instant on the site's quick streaming stream of tweets. To evade points that are famous routinely Twitter centers around subjects that are being talked about considerably more than expected themes that as of late endured an expansion of utilization, so it.

The file is encrypted using the hash value as a key. The hash value is then encrypted using the public keys of authorized readers and attached to the file as metadata. Convergent encryption ensures that identical encrypted files are recognized as identical. However, there is still the challenge of identifying these files across a large number of machines in a robust and decentralized manner. A cloud user is someone who wants to outsource data on public storage, which functions as a public cloud in cloud computing. The system provides authentication for users to enter and upload data with specific privileges for accessing the uploaded data. Public storage refers to a storage disk that allows users to store their data on it, with authorization to prevent duplicate data from being uploaded.

ADVANTAGES OF PROPOSED SYSTEM:

- we are convinced that client side deduplication is more efficient in the term of network bandwidth saving.
- In the deduplication system the encryption technique plays vital role to ensure security of the user throughout the system.

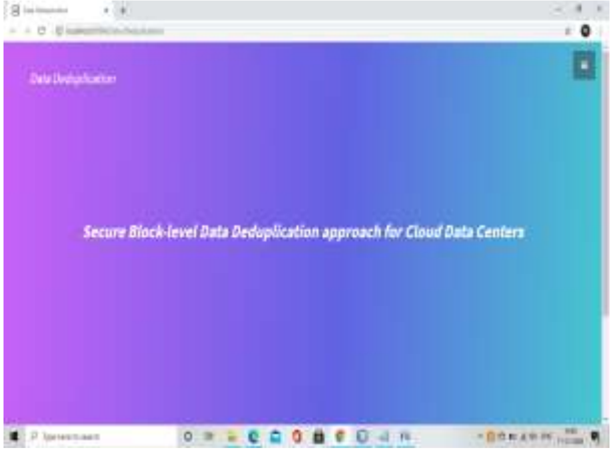
MODULES:

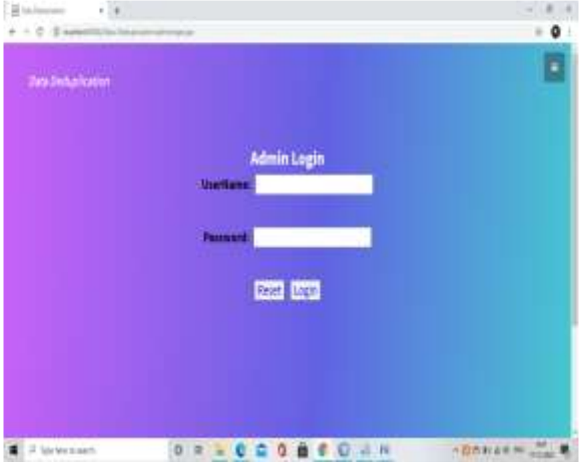
- User
- Admin

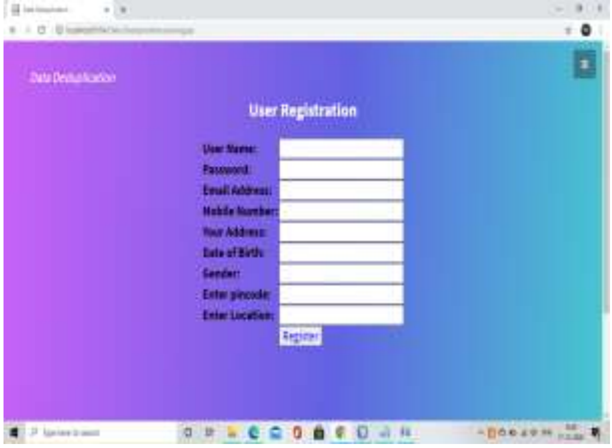
User: User login valid id and password after that user upload the file to cloud and view the file .if user want edit the file he can edit file and store to the cloud and user download the file by using mac key and verify the key after that download file.

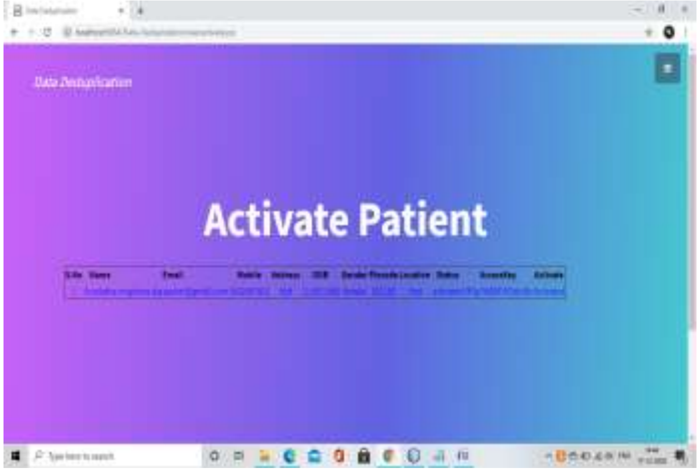
Admin: Admin login the valid id and password after activate the user and maintain the details to the users and maintain download history.

SAMPLE SCREENSHOTS


Data Deduplication
Secure Block-level Data Deduplication approach for Cloud Data Centers

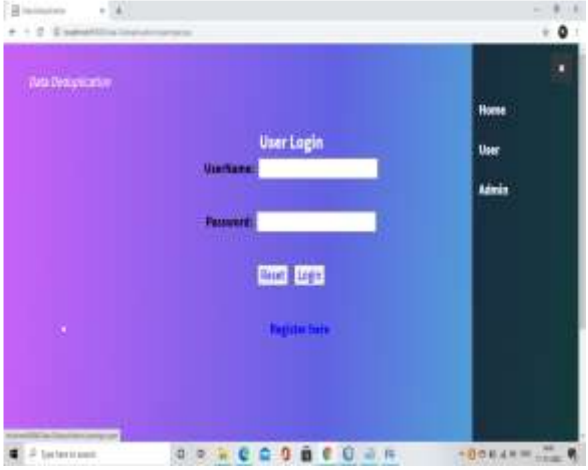


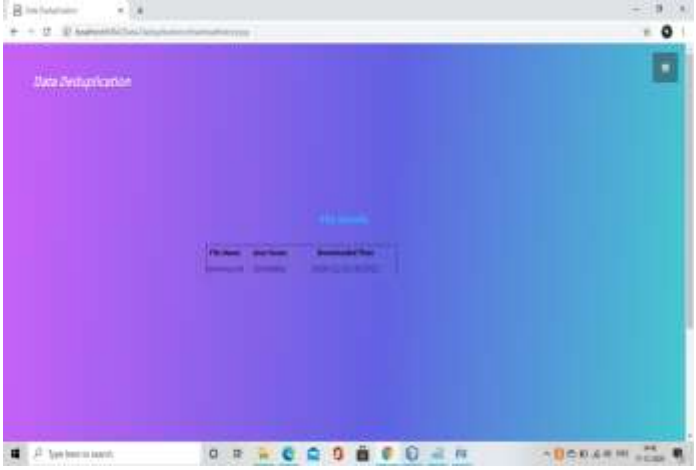




Activate Patient

S.No	Name	Email	Mobile	Address	DOB	Gender	Pincode	Location	Status	Secretkey	Activate
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]





File Name	User Name	Downloaded Time
[illegible]	[illegible]	[illegible]

III. CONCLUSION:

In this paper, we proposed a secure medical diagnosis and treatment framework named as P-Med that can be used to recommend therapy methods to the patients according to their illness states. The medical model in P-Med is constructed based on NFA, encrypted and outsourced to cloud. The patient submits successive several days of encrypted mIoT data to issue a query and get the top-k best treatment recommendations using secure selection algorithm. A secure illness state match protocol is also designed in P-Med to achieve quantitative secure comparison between the state in medical model and patient's illness state that are monitored by mIoT.Moreover, secure NFA evaluation method in P-Med reduces the interaction between cloud and patient to a single round. Finally, we evaluate the security and performance of P-Med moving through twitter. Understanding twitter was as significant as knowing the subjects being referred to. The consequences of the past investigations, driven us to the end that highlight choice is a totally need in a content grouping framework. This was demonstrated when we contrasted our outcomes and a framework that utilizes precisely the same dataset.

VI. REFERENCES

[1] Young K, Gupta A, Palacios R. Impact of telemedicine in pediatric postoperative care. Telemedicine and e-Health. 2018 Dec 5.

[2] Verma P, Sood S K. Cloud-centric IoT based disease diagnosis healthcare framework[J]. Journal of Parallel and Distributed Computing, 2018, 116:27-38.

[3] Kumar P M, Lokesh S, Varatharajan R, et al. Cloud and IoT based disease prediction and diagnosis system for healthcare using Fuzzy neural classifier[J]. Future Generation Computer Systems, 2018, 86: 527-534.

[4] Sipser M. Introduction to the theory of computation (3rd Edition). Cengage Learning (2013).

[5] Gambheer H. Design safety verification of medical device models using automata theory[D]. California State University Channel Islands, 2016.

[6] Alkhaldi F, Alouani A. Systemic design approach to a real-time healthcare monitoring system: reducing unplanned hospital readmissions[J]. Sensors, 2018, 18(8): 2531.

[7] Caballero-Ruiz E, et al. A web-based clinical decision support system for gestational diabetes: Automatic diet prescription and detection of insulin needs[J]. International Journal of Medical Informatics, 2017, 102: 35-49.

[8] Sasakawa H, Harada H, Duverle D, et al. Oblivious evaluation of nondeterministic finite automata with application to privacy-preserving virus genome detection[C]. WPES 2014:21-30, ACM.

[9] Papageorgiou A, etc. Security and privacy analysis of mobile health applications[J]. IEEE Access. 2018(6):9390-403.

[10] Droste M, Kuich W, Vogler H, editors. Handbook of weighted automata. Springer Science & Business Media, 2009, Sep 18.

[11] Paillier P. Public-key cryptosystems based on composite degree residuosity classes[C]. Eurocrypt 1999: 223-238, Springer.

[12] Bresson E, Catalano D, Pointcheval D. A simple public-key cryptosystem with a double trapdoor decryption mechanism and its applications[C]. ASIACRYPT 2003: 37-54, Springer.

[13] Yang Y, Liu X, Deng R. Multi-user multi-keyword rank search over encrypted data in arbitrary

language[J]. IEEE Transactions on Dependable and Secure Computing, 2017, DOI: 10.1109/TDSC.2017.2787588.

[14] Liu X, Deng R, Choo K K R, et al. An efficient privacy-preserving outsourced calculation toolkit with multiple keys[J]. IEEE Transactions on Information Forensics and Security, 2016,11(11):2401.

[15] Yang Y, Liu X, Deng R. Expressive query over outsourced encrypted data[J]. Information Sciences, 2018, 442: 33-53.al Image Analysis." In CLOSER, pp. 350-357. 2013