iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 6 **Issue:** II **Month of publication:** February 2018

**DOI:**

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor : 6.887*
*Volume 6 Issue II, February 2018- Available at www.ijraset.com*

# A Secure Access Policies Based Data Deduplication System

Mr. T. A. MohanaPrakash[1], Nivas Prabu V [2], Vigneshvaran G [3] , Prince E. T [4],
[1]*Associate Professor*
[2, 3, 4]*Students of B.E. Computer Science Department* [2,3,4]*Panimalar Institute of Technology, Chennai, Tamil Nadu, India.*

***Abstract:*** ***Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. Big Data in most companies are processed by Hadoop by submitting the jobs to Master. The Master distributes the job to its cluster and process map and reduces tasks sequentially. But nowadays the growing data need and the competition between Service Providers leads to the increased submission of jobs to the Master. This Concurrent job submission on Hadoop forces us to do Scheduling on Hadoop Cluster so that the response time will be acceptable for each job. In this Deduplication techniques are most widely employed to backup data and minimize network and storage overhead by detecting and eliminating redundancy among data. So which is crucial for eliminating duplicate copies of identical data in order to save storage space and network bandwidth? We present an attribute-based storage system with secure deduplication in a hybrid cloud setting, using public cloud and private cloud. Where a private cloud is responsible for duplicate detection and a public cloud manages the storage. Instead of keeping multiple data copies with the same content, the system eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Each such copy can be defined based on user access policies. In this user will upload the file with access policies and then file type question with answer. Then same file with different access policies to set the particular file to replace the reference. Where a user's private key is associated with an attribute set, a message is encrypted under an access policy over a set of attributes, and a user can decrypt a ciphertext with his/her private key if his/her set of attributes satisfies the access policy associated with this ciphertex***

## I. RELATED WORK

In this paper, we present an attribute-based storage system which employs ciphertext-policy attribute-based encryption (CP-ABE) and supports secure deduplication. To enable the deduplication and distributed storage of the data across HDFS. And then using two way cloud in our storage system is built under a hybrid cloud architecture, where a private cloud manipulates the computation and a public cloud manages the storage. The private cloud is provided with a trapdoor key associated with the corresponding ciphertext, with which it can transfer the ciphertext over one access policy into ciphertexts of the same plaintext under any other access policies without being aware of the underlying plaintext. After receiving a storage request, the private cloud first checks the validity of the uploaded item through the attached proof. If the proof is valid, the private cloud runs a tag matching algorithm to see whether the same data underlying the ciphertext has been stored. If so, whenever it is necessary, it regenerates the ciphertext into a ciphertext of the same plaintext over an access policy which is the union set of both access policies. like public cloud and private cloud. We have shown the concept of deduplication effectively and security is achieved by means of Proof of Ownership of the file. That is attribute-based storage system ciphertext-policy attribute-based encryption (CP-ABE) and supports secure deduplication

## II. LITERATURE SURVEY

| S.NO | TITLE | AUTHOR | CONCEPT | YEAR | ADVANTAGE | DISADVANTAGE |
|---|---|---|---|---|---|---|
| 1 | Cloud based Storage Drive Forensics | Prashant Bhatt1 | Cloud Storage is recently as emerging topic in these eras. As the data are increasing, the storage become major issue for the people. There are different kind of Cloud Storage application such as One Drive, Sky Drive, Drop | 2004 | On Agile methods, an experiment that test-first nature of TDD and compared it to the test-last nature of traditional software processes. | This retrieves only outline the contributions of this research for understanding of TDD. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | Box and Google Drive. Google Drive is gaining more popularity as it is user friendly than any other Cloud Storage Application. Google Drive is a Cloud Storage Application which allows user to store, share and edit the file in the cloud. | | | |
| 2 | Fuzzy Identity Based Encryption | Amit Sahai. | The project introduces a new type of Identity Based Encryption (IBE) scheme that we call Fuzzy Identity Based Encryption. A Fuzzy IBE scheme allows for a private key for an identity id to decrypt a ciphertext encrypted with another identity id0 if and only if the identities id and id0 are close to each other as measured by some metric (e.g. Hamming distance). A Fuzzy IBE scheme can be applied to enable encryption using biometric measurements as identities. | 2010 | The system is able to detect imitation attacks, allow for the encryption of data using a biometric input. | The inherent non-determinism makes it difficult to extract a cryptographic key from a biometric input. It does not fit into the paradigm of Identity Based Encryption. |
| 3 | Message-Locked Encryption for Lock-Dependent Messages | Martin Abadi | Motivated by the problem of avoiding duplication in storage systems, Bellare, Keelveedhi, and Ristenpart have recently put forward the notion of Message-Locked Encryption (MLE) schemes which sub-sumes convergent encryption and its variants. Such schemes do not rely on permanent secret keys, but rather encrypt messages using keys de- rived from the messages themselves. We strengthen the notions of security proposed by Bellare et al. by con- sidering plaintext distributions that may depend on the public parame- ters of the schemes. We refer to such inputs as lock-dependent messages. We construct two schemes that satisfy our new notions of | 2014 | The system is fully randomized scheme. The inputs are lock-dependent messages. | The system uses computationally expensive NIZKs to identify all duplicate ciphertexts. The identical plaintexts are mapped only to identical ciphertexts. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | security for message-locked encryption with lock-dependent messages. | | | |
| 4 | Message-Locked Encryption and Secure Deduplicati on | Mihir Bellar | The system formalizes a new cryptographic primitive, Message-Locked Encryption (MLE), where the key under which encryption and decryption are performed is itself derived from the message. MLE provides a way to achieve secure deduplication (space-e cient secure outsourced storage), a goal currently targeted by numerous cloud-storage providers. We provide de nitions both for privacy and for a form of integrity that we call tag consistency. Based on this foundation, we make both practical and theoretical contributions. | 2008 | It has the improvement in fuel economy using DP based charge-depletion control compared to rule based control. | On the traffic flow on highway with on/off ramps which may be missed by the model which used for only main road detectors data. |
| 5 | Digital Forensic Trends and Future | Farhood Norouzizadeh Dezfoli, Ali Dehghantanha, Ramlan Mahmoud, Nor Fazlida Binti Mohd Sani, Farid Daryabar | Nowadays, rapid evolution of computers and mobile phones has caused these devices to be used in criminal activities. Providing appropriate and sufficient security measures is a difficult job due to complexity of devices which makes investigating crimes involving these devices even harder. Digital forensic is the procedure of investigating computer crimes in the cyber world. Many researches have been done in this area to help forensic investigation to resolve existing challenges. This paper attempts to look into trends of applications of digital forensics and security at hand in various aspects and provide some estimations about future research trends in this area. | 2010 | It maintains the privacy of the clients.<br>It helps in exploiting security mechanism and framework rather than privacy protection techniques. | It is not so responsive to the current security trends and issues. The leakage of the issue may attract huge media attention resulting in endangering the reputation. |
| 6 | Survey on | Renjith P , | Cloud storage helps | 2013 | It is a simple | No access control |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Data Sharing and Re-Encryption in Cloud | Sabitha S | enterprises and government agencies significantly reduce their financial overhead of data management, since they can now archive their data backups remotely to third-party cloud storage providers rather than maintain data centers of their own. Security concerns become relevant as we outsource the storage of possibly sensitive data to third party cloud storage. Data stored in cloud may be unexpectedly disclosed in the future due to malicious attacks on the cloud or careless management of cloud operators. Secure data transfer is needed to maintain the data security between authorized users. The challenge of achieving secure data sharing is that we have to encrypt the data and at the same time it should be available to those authorized clients. This is made possible through re-encryption. | | method of data sharing that does not use either ABE or time based encryption. It directly stores data and keys in servers as shares. | mechanisms are implemented to handle revoked users. Owner signature is embedded in the data. |
| 7 | Forensic Analysis of Google Drive on Windows | Ming Sang Chang | Cloud storage services are increasingly used by consumers, business, and government. These services are fairly easy to obtain. Google Drive is a popular service, providing users a cost-effective, and in some cases free, ability to access, store, collaborate, and disseminate data. It is difficult to identify, acquire, and preserve the evidences when criminals use disparate services. This study was undertaken to determine the data remnants on a Windows computer. | 2015 | The remnants of cloud activity can be found on local machines. It could be valuable for the forensic examiners. | This may enable investigators to identify the location of data. In this research, an investigator can identify Google Drive account use by undertaking keyword searches and examine test files locations to locate relevant information. |
| 8 | Proxy Re-encryption Schemes for | Raghi Roy, Paul P. Mathai, | This paper presents a survey on Proxy re encryption techniques | 2015 | PRE has captured a lot of concern due to the delegation | Encoding operations over encrypted |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Secure Cloud Data And Applications: A Survey | | with respect to secure cloud data and its application. To keep sensitive user data confidential against untrusted servers, cryptographic methods are used to provide security and access control in clouds. As the data is shared over the network, it is needed to be encrypted. There are many encryption schemes that provide security and access control over the network.Proxy re-encryption enables the semi-trusted proxy server to re-encrypt the ciphertext encrypted under Alice's public key to another ciphertext encrypted under Bob's public key. The re-encryption is done without the server being able to decrypt the ciphertext. | | function of decryption. PRE is also an essential technique as many real time applications. | messages is not possible. The key privacy proof is more difficult than that of CPA security |
| 9 | Efficient Deduplication in Disk- and RAM-based Data Storage Systems | Andrej Tolic , Andrej Brodnik. | Modern storage systems such as distributed file systems and key-value stores in many cases exhibit data redundancy. The issue is addressed by deduplication, a process of identifying and eliminating duplicate data. While deduplication is typically applied to data stored on disks, the emergence of RAM-based storage systems opens new problems on one hand while being insensitive to some inherent deficiencies of deduplication such as fragmentation. In this paper we present a review of disk- and memory-based deduplication. | 2009 | Most of the techniques in deduplication were developed with backup/ archival or virtualization workloads in mind. We notice that work on memory data deduplication is scarce and mostly focused on virtualization. | Subpage-level memory deduplication. Existing systems only deduplicate whole pages. Better understanding the effect of different storage APIs on deduplication. |
| 10 | Avoiding | Benjamin Zhu, | Disk-based deduplication | 2012 | The techniques for | These techniques |

| | | | | | | |
|---|---|---|---|---|---|---|
| | the Disk Bottleneck in the Data Domain Deduplicati on File System | Kai Li, Hugo Patterson | storage has emerged as the new-generation storage system for enterprise data protection to replace tape libraries. Deduplication removes redundant data segments to compress data into a highly compact form and makes it economical to store backups on disk instead of tape. A crucial requirement for enterprise data protection is high throughput, typically over 100 MB/sec, which enables backups to complete quickly. A significant challenge is to identify and eliminate duplicate data segments at this rate on a low-cost system that cannot afford enough RAM to store an index of the stored segments and may be forced to access an on-disk index for every input segment. | | minimizing disk I/Os to achieve good deduplication performance match well against the industry trend of building many-core processors. | are general methods to improve throughput performance of deduplication storage systems. |

## III. CONCLUSION

In this project, the new distributed de-duplication systems with file-level and fine-grained block-level data deduplication, higher reliability in which the data chunks are distributed across HDFS storage, reliable key management in secure de-duplication and the security of tag consistency and integrity were achieved.

## IV. FUTURE WORK

With the goal of saving storage space for cloud storage services, first solution for balancing confidentiality and efficiency in performing deduplication called attribute-based encryption was proposed, where a message is encrypted under a message-derived key so that identical plaintexts are encrypted to the same cipher texts. In this case, if two users upload the same file, the cloud server can discern the equal cipher texts and store only one copy of them.

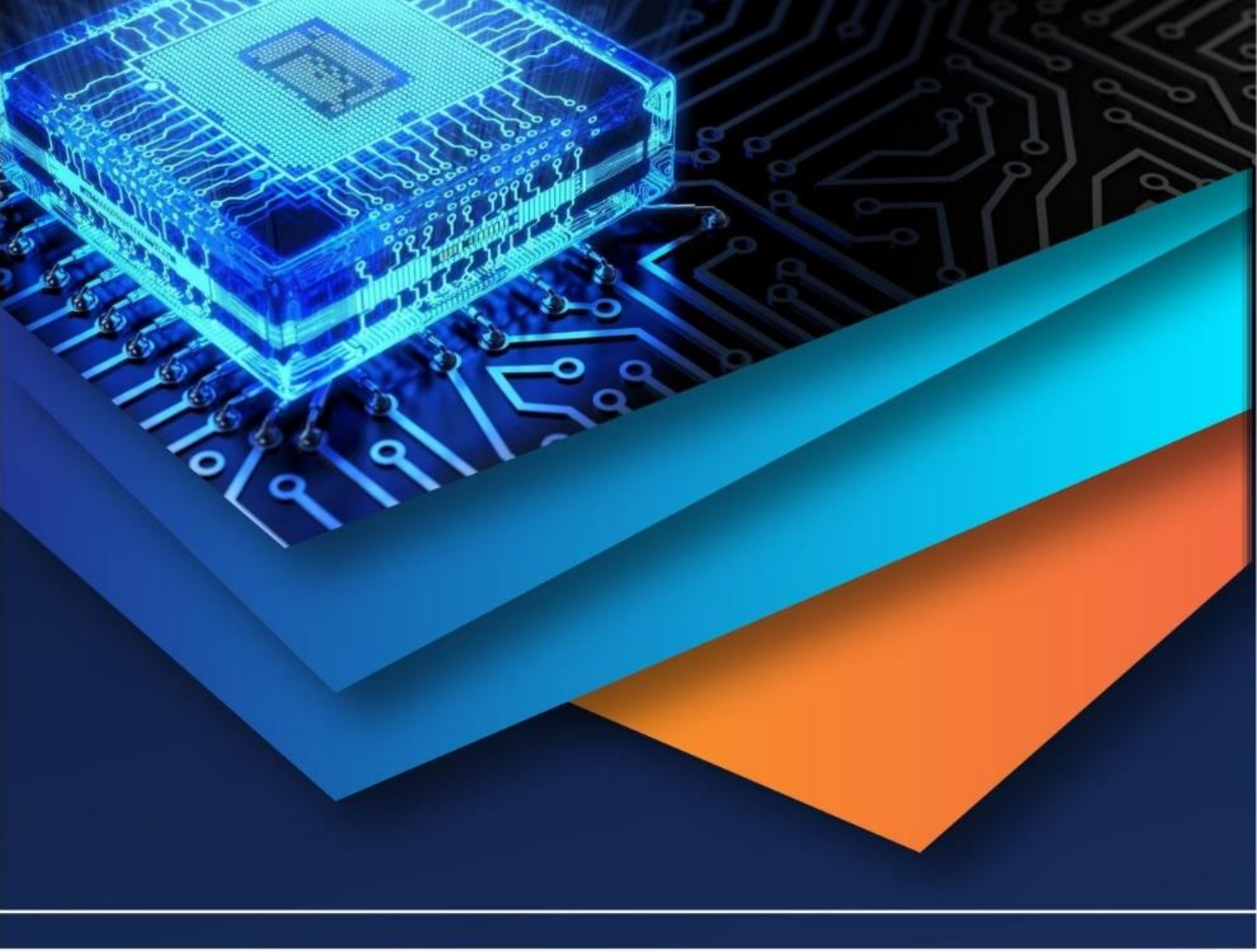
## REFERENCES

[1] D. Quick, B. Martini, and K. R. Choo, Cloud Storage Forensics. Syngress Publishing / Elsevier, 2014. [Online]. Available: http://www.elsevier.com/books/cloud-storageforensics/ quick/978-0-12-419970-5
[2] K. R. Choo, J. Domingo-Ferrer, and L. Zhang, “Cloud cryptography: Theory, practice and future research directions,” Future Generation Comp. Syst., vol. 62, pp. 51–53, 2016.
[3] K. R. Choo, M. Herman, M. Iorga, and B. Martini, “Cloud forensics: State-of-the-art and future directions,” Digital Investigation, vol. 18, pp. 77–78, 2016.
[4] Y. Yang, H. Zhu, H. Lu, J.Weng, Y. Zhang, and K. R. Choo, “Cloud based data sharing with fine-grained proxy re-encryption,” Pervasive and Mobile Computing, vol. 28, pp. 122–134, 2016.
[5] D. Quick and K. R. Choo, “Google drive: Forensic analysis of data remnants,” J. Network and Computer Applications, vol. 40, pp. 179–193, 2014.
[6] A. Sahai and B. Waters, “Fuzzy identity-based encryption,” in Advances in Cryptology - EUROCRYPT 2005, 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005, Proceedings, ser. Lecture Notesin Computer Science, vol. 3494. Springer, 2005, pp. 457–473.
[7] B. Zhu, K. Li, and R. H. Patterson, “Avoiding the disk bottleneck in the data domain deduplication file system,” in 6th USENIX Conference on File and Storage Technologies, FAST 2008, February 26-29, 2008, San Jose, CA, USA. USENIX, 2008, pp. 269–282.

[8] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Advances in Cryptology - EUROCRYPT 2013, 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece,May 26-30, 2013. Proceedings, ser. Lecture Notes in Computer Science, vol. 7881. Springer, 2013, pp. 296–312.

[9] M. Abadi, D. Boneh, I. Mironov, A. Raghunathan, and G. Segev, "Message-locked encryption for lock-dependent messages," in Advances in Cryptology - CRYPTO 2013 - 33rd Annual Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2013. Proceedings, Part I, ser. Lecture Notes in Computer Science, vol. 8042. Springer, 2013, pp. 374–391.

[10] S. Keelveedhi, M. Bellare, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in Proceedings of the 22th USENIX Security Symposium, Washington, DC, USA, August 14-16, 2013. USENIX Association, 2013, pp. 179–194.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY