Effective Scheme for data deduplication using secret key sharing in the cloud environment

M.chennakesavarao[1], A. Alekhya[2]

[1]*Assoc. Prof, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India*

[2]*PG Scholar, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India*

ABSTRACT - Data deduplication has become a vital aspect of managing repositories of outsourced data to the cloud data centers. However, centralized data centers suffer data loss and accessibility difficulties if something goes faulty, as deduplication retains a unique content copy. Secure data deduplication utilizes convergent encryption to perform data deduplication in the encrypted domain. However, handling convergent keys offers a single point of vulnerability and overhead difficulties. Towards this end, we propose a secure data deduplication scheme that formally addresses fault tolerance, efficient and reliable key management, data confidentiality by obfuscation of outsourced information, an integrity check at the user's end before downloading via computation of authentication codes. Data is distributed into random-seeming shares based on the Permutation ordered binary (POB) number scheme at many servers and is further made safe using the notion of proof of ownership (PoW) concept. Also, key overhead is decreased utilizing Chinese Remainder Theorem (CRT) based secret sharing. The experimental findings have proved the efficacy of the suggested technique, and security analysis supports its applicability concerning various threats in real-time scenarios.

Keywords: Cloud, deduplication, security

I. INTRODUCTION

Cloud computing is a type of Internet-based hardware and software solution made available via a third-party service provider. It was named after the cloud-shaped symbol commonly used to represent the system's intricate framework in system diagrams. With cloud computing, an individual's data, software, and computational abilities are entrusted to remote services. High-end software and server networks are standard features of these services.

- A significant goal of cloud computing is to give military-grade computing power to the private sector to conduct tens of trillions of computations per second, allowing individuals to run complex financial portfolios, get personalized information, or save data.
- The NIST's definitions of cloud computing are paraphrased as follows:
- Just-in-time self-service: A user can select computing resources to be made available on an as-needed basis without consulting each service's supplier.
- Good network connectivity: Capabilities are accessible through the network, and they are used by a variety of different client types, including thin and thick clients (e.g., mobile phones, laptops, and PDAs).
- This means that physical and virtual resources are dynamically assigned and reassigned according to the needs of individual customers, all within a multi-tenant architecture. Though the consumer has no information or control over the location of the resources, they may still identify the location in broad categories (e.g., country, state, or datacenter). Storage, processing, memory, network bandwidth, and virtual machines are all examples of resources.
- Capabilities can be fast and flexibly provided in some circumstances, automating deployment and rollout. For the customer, it looks like there's no limit to the capabilities available to provision. They can be bought in any quantity, at any time.
- The service's measured utility: Cloud systems offer an auto-control and optimization system for the resources, using a metering capability at some level of abstraction suited to the service's kind (e.g., storage, processing, bandwidth, and active user accounts). Manage, control, and report resource utilization to ensure transparency for both the provider and the service consumer.

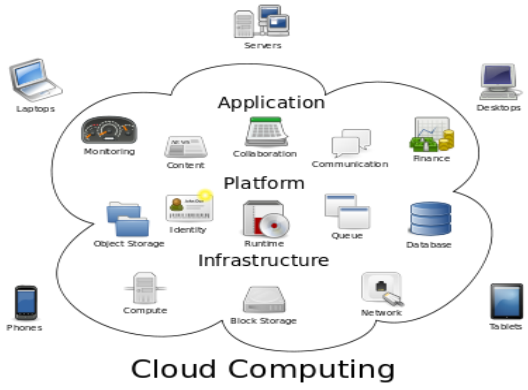

Figure 1. Structure of cloud computing

II. RELATED WORK

Cloud-based architecture is becoming widely embraced, but one of the most pressing concerns is keeping content private. While some research has addressed this goal with classic encryption methods, such as [15], there have been many difficulties. Although the security was assured, the size of the encrypted content increased because each user used their key to encrypt it. Because of this, the desire for better storage options was prominent.

Additionally, holding on to the encryption keys for users proved to be burdensome as the amount of data grew. If any transmission errors caused the keys to getting corrupted, the encrypted content could not be retrieved. In addition, the keys provided a single point of vulnerability, and hence they may constitute a security risk to the system if they were compromised.

A threshold-based secret sharing mechanism has been utilized to incorporate fault tolerance in this regard (16). To combat the rising expense of storage and manage the growing quantity of encrypted content, convergent encryption's use of deduplication has become a solution, as shown in Fig. 1. A content-derived key in these systems encrypts the content. So, the plaintexts being identical would result in the same ciphertext [10]. This can be used to encrypt and deduplicate the content while still keeping it private. Bellare et al. [5] have codified the concept. Several contemporary encryption techniques have been developed to address the storage issues that plague cloud service providers, but critical management remains a difficulty [3].

The PoW (Proof of Work) method has been suggested to keep communication costs in deduplication methods under control. In place of uploading files repeatedly, the client must verify their identity to the server. The Merkle Hash Tree-based Proof of Work method activated the bounded leakage option [14].

A different PoW-based approach involved random sampling of bit locations [9]. Even though a large number of Proof of Work (PoW) systems have been presented, few can handle encrypted files [20].

III. PROPOSED ARCHITECTURE

For apps to be wise, they need a speedy response and processing vast amounts of data, just like the Internet of Things (IoT). To achieve both of these at once, multi-level data processing is needed, as shown in Figure 2. The time-sensitive data gathering and filtering procedures in the first order are a potential resource from Rim computing. After all, the most effective resource may be crucial to more complex operations.

This method of protection is through making something hard to understand. We protect information on the cloud by screening which has access and identifying abnormal access patterns. If we find that someone is accessing our database without permission, we email them a list of targets to distract them. This prevents consumers' knowledge from being misused.

This is enhanced in the cloud download age. One way to obtain information about an unauthorized user's activity is to produce distraction data, for example, decoy documents, nectar documents, nectar pots, and other falsified material. If your intruder is not operating with beneficial, current info, serving imitations can make them hop and confuse. This new development can work in tandem with customer innovation profiling to protect consumer data in the cloud. Anywhere inconsistencies in cloud access are spotted, cloud-based data about distractions can be delivered and transmitted to look entirely familiar and realistic.

Fog computing (Fog) is a network that employs client-end devices to create multiple processing, control, configuration, measurement, and management resources from storage, communication, and computing resources. Several different disciplines have worked on it since its beginning. Some consider fog computing an offshoot of edge computing or equal to it, although most consider it a supplement to cloud computing. As a result, it's not clear what it means or how to define it.

IV. RESULTS AND OBSERVATION

Many cloud-based services show how a user can connect to the Internet and store data, files, and media in a remote server. To safeguard user data, ensure that the user has access to it while limiting others' access. Classification protection remains one of the leading security challenges that many people dislike.

Standard access restrictions and encryption were suggested to secure data in the cloud, as many proposals were put forth. While specific techniques have become flawed due to internal breaches, broken services, and faulty implementation, the list of reasons older methods have fallen short of expectations continues to grow. Flawed design, failure to detect inventive attack methods, and a range of other problems also contribute to the overall problems of the older methods [8]. It is impossible to restore an atmosphere where people can have faith in cloud computing due to the recurrence of attacks and their eroding knowledge. It's best to prepare oneself for situations like this.

A basic rule of thumb is that if we make the stolen information less valuable to the thief, we can cut down on the damage done to the stolen data. This is something that can be done with a preemptive strike against misinformation. We advocate that cloud-based services can be developed more stably if we integrate two extra security options.

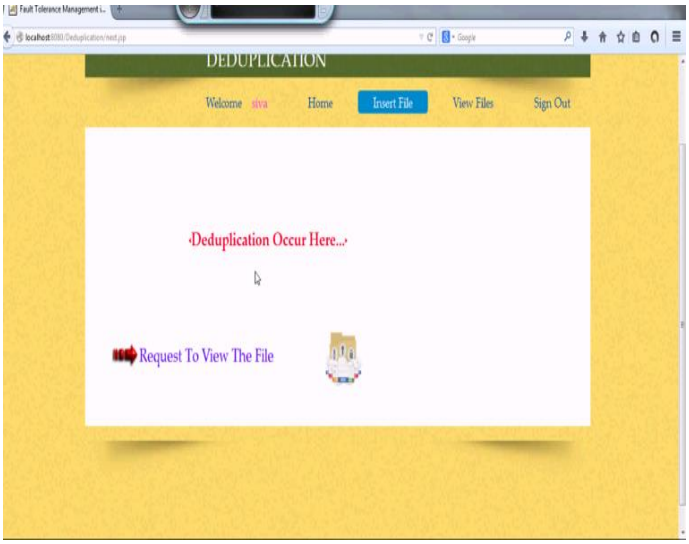

Figure 3. Results proposed system

V. CONCLUSION

This article offers a novel method to secure personal and corporate data in the cloud. We recommend profiling user behavior to monitor how people access documents in a cloud service to discover if an untrustworthy insider accesses documents inappropriately. Documents stored in the cloud containing dummy data are also employed as sensors to detect illicit access. To flush out any malevolent insiders who are likely to compromise sensitive information, we deluge these people with fake information. Social networks and the cloud could enjoy unparalleled levels of security thanks to these types of pre-emptive attacks that rely on misinformation technologies.

VI. REFERENCES

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," PLoS ONE, vol. 13, no. 3, Mar. 2018, Art. No. e0194889.

[2] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in Proc. Eur. Bus. Intell. Summer School. Berlin, Germany: Springer, 2012, pp. 62–77.

[3] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, "Regression models for prognostic prediction: advantages, problems, and suggested solutions," Cancer Treat. Rep., vol. 69, no. 10, pp. 1071–1077, 1985

[4] P. Lapuerta, S. P. Azen, and L. Labree, "Use of neural networks in predicting the risk of coronary artery disease," Comput. Biomed. Res., vol. 28, no. 1, pp. 38–52, Feb. 1995.

[5] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B.Kannel, "Cardiovascular disease risk profiles," Amer. heart J., vol. 121, no. 1, pp. 293–298, 1991.

[6] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Comput. Sci., vol. 83, pp. 1064–1069, Jan. 2016.

[7] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19," PLoS ONE, vol. 15, no. 3, Mar. 2020, Art. No. e0231236.

[8] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast during an emergency response," JAMA, vol. 323, no. 16, p. 1545, Apr. 2020.

[9] WHO. Naming the Coronavirus Disease (Covid-19) and the Virus That Causes it. Accessed: Apr. 1, 2020. [Online]. Available: https:// www.who.int/emergencies/diseases/novel-coronavirus-2019/technic% alguidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virusthat-cau%ses-it

[10] C. P. E. R. E. Novel, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (Covid-19) in China," Zhonghua Liu Xing Bing Xue Za Zhi= Zhonghua Liuxingbingxue Zazhi, vol. 41, no. 2, p. 145, 2020.

[11] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, and B. Berkhout, ‘‘Identification of a new human Coronavirus,’’ Nature Med., vol. 10, no. 4, pp. 368–373, 2004.

[12] Johns Hopkins University Data Repository. Cssegisanddata. Accessed: March 27, 2020. [Online]. Available: https://github.com/ CSSEGISandData

[13] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, "Analytics defined," in Information Security Analytics, M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, Eds. Boston, MA, USA: Syngress, 2015, pp. 1–12. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128002070000010

[14] H.-L. Hwa, W.-H. Kuo, L.-Y. Chang, M.-Y. Wang, T.-H. Tung, K.-J. Chang, and F.-J. Hsieh, "Prediction of breast cancer and lymph node metastatic status with tumor markers using logistic regression models," J. Eval. Clin. Pract., vol. 14, no. 2, pp. 275–280, Apr. 2008.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," J. Roy. Stat. Soc., Ser. B, Methodol., vol. 58, no. 1, pp. 267–288, Jan. 1996.

[16] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[17] X. F. Du, S. C. H. Leung, J. L. Zhang, and K. K. Lai, "Demand forecasting of perishable farm products using support vector machine," Int. J. Syst. Sci., vol. 44, no. 3, pp. 556–567, Mar. 2013.

[18] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, "Tweets classification on the base of sentiments for US

airline companies," Entropy, vol. 21, no. 11, p. 1078, Nov. 2019.

[19] E. Cadenas, O. A. Jaramillo, and W. Rivera, "Analysis and forecasting of wind velocity in Chetumal, Quintana roo, using the single exponential smoothing method," Renew. Energy, vol. 35, no. 5, pp. 925–930, May 2010.

[20] J. Lupón, H. K. Gaggin, M. de Antonio, M. Domingo, A. Galán, E. Zamora, J. Vila, J. Peñafiel, A. Urrutia, E. Ferrer, N. Vallejo, J. L. Januzzi, and A. Bayes-Genis, ''Biomarker-assist score for reverse remodeling prediction in heart failure: The ST2-R2 score,'' Int. J. Cardiol., vol. 184, pp. 337–343, Apr. 2015.

[21] J.-H. Han and S.-Y. Chi, "Consideration of manufacturing data to apply machine learning methods for predictive manufacturing," in Proc. 8th Int. Conf. Ubiquitous Future Netw. (ICUFN), Jul. 2016, pp. 109–113.

[22] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Climate Res., vol. 30, no. 1, pp. 79–82, 2005.

[23] R. Kaundal, A. S. Kapoor, and G. P. Raghava, "Machine learning techniques in disease forecasting: A case study on rice blast prediction," BMC Bioinf., vol. 7, no. 1, p. 485, 2006.

[24] S. Baran and D. Nemoda, "Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting," Environmetrics, vol. 27, no. 5, pp. 280–292, Aug. 2016.

[25] Y. Grushka-Cockayne and V. R. R. Jose, "Combining prediction intervals in the m4 competition," Int. J. Forecasting, vol. 36, no. 1, pp. 178–185, Jan. 2020.

[26] N. C. Mediaite. Harvard Professor Sounds Alarmed on 'Likely' Coronavirus Pandemic: 40% to 70% of world Could be Infected This Year. Accessed: February 18, 2020. [Online]. Available: https://www.mediaite.com/news/ Harvard-professor-sounds-alarm-on-likely-%coronavirus-pandemic-40- to-70-of-world-could-be-infected-this-year

[27] BBC. Coronavirus: Up to 70% of Germany Could Become Infected— Merkel. Accessed: Mar. 15, 2020. [Online]. Available: https://www.bbc. com/news/world-us-canada-51835856