Data Duplication Removal and Image, Text Deduplication Using ML

Mrs. Kavyashree H N,
Assistant professor
Department Of Computer Engineering
Nutan Maharashtra Institute of
Engineering and Technology, Talegaon, Pune.

Mr. Himanshu T. Sarode
Department Of Computer Engineering
Nutan Maharashtra Institute of
Engineering and Technology, Talegaon, Pune.

Mr. Yash S. Pawar
Department Of Computer Engineering
Nutan Maharashtra Institute of
Engineering and Technology, Talegaon, Pune.

Mr. Mohanish K. Sarode
Department Of Computer Engineering
Nutan Maharashtra Institute of
Engineering and Technology, Talegaon, Pune.

Abstract: In the era of information development, information transfer is far more practical. However, there is always a chance that the transmission method may be compromised, stolen, or assaulted, raising questions about the accuracy of the data source. Several academics suggested using passwords to secure sensitive data because of this. Proposed 3D-Playfair Cipher with Message Integrity Using MD5. The encryption used in this study is 3D-Playfair. Nevertheless, basic 3D-playfair encryption is unable to ensure the integrity of data while it is being sent, thus the author suggests In addition to using MD5 to guarantee data integrity, this study employs XOR calculation techniques to further confirm the data's legitimacy because there are concerns over the reliability of the data source.

Keywords: Data security, Deduplication, Authorization, Authentication, Access control, Cloud Security, Cipher Technology, Data synchronization.

I. INTRODUCTION

Large volumes of data are gathered by businesses from a variety of industries in today's data-driven world in order to support their operations, strategic objectives, and decision-making procedures. One recurring issue, though, is duplicate data in these databases. In addition to taking up expensive storage space, duplicate data taints data analysis, creates inconsistencies, and threatens the dependability of electronic systems [9]. Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied.

Through the use of a strong deduplication technique, this research seeks to address the problem of data duplication. Deserialization, sometimes referred to as duplicate finding or duplicate removal, is locating and eliminating duplicated entries from datasets [10]. Organizations may increase the effectiveness of data-driven operations, maintain data quality, and streamline procedures by getting rid of duplicates. Creating and implementing an effective deduplication system that can locate and eliminate duplicate entries in massive datasets is the main goal of this project. To effectively identify duplicates while reducing false positives and maintaining data integrity, this system will make use of sophisticated algorithms and methodologies. Additionally, the goal of this research is to investigate several facets of deduplication, such as algorithm choice, scalability, performance enhancement, and interaction with current data management processes. Through an all-encompassing approach, the suggested solution seeks to offer a complete method for managing data duplication. With the help of this project, we hope to improve data quality management techniques and enable businesses to fully utilize their data assets. Organizations may improve decision-making, stimulate creativity, and increase operational efficiency by reducing the effects of data duplication

In an environment where data explosion is commonplace, organizations struggle to effectively manage enormous amounts of data. In the middle of this flood, redundant data becomes a powerful enemy that impedes decision-making, data analysis, and resource allocation. When faced with diverse data sources and dynamic content formats, traditional deduplication techniques frequently break down [11]. Our project sets out to pioneer a unified solution for deduplication across textual and visual domains by delving into the field of machine learning in order to address this problem.

II. LITERATURE SURVEY

A secondary encryption secure transmission strategy to safeguard the original private information while offering quality-of-service provisioning for secondary system [1]. The suggested technique encrypts the principal confidential messages using the secure secondary messages, allowing the secondary system to get certain spectrum opportunities. To be more precise, in cases where the primary system is secure, the primary information can be transmitted directly; in cases where the primary system is not secure but the secondary messages can be transmitted securely, the primary system encrypts the primary information using the secure secondary messages; in the absence of this, the spectrum will be used for secondary transmission. We examine how well the average secondary throughput and the primary ergodic secrecy rate perform for the suggested architecture [2].

Due to its numerous uses in a variety of fields, including cloud computing and the transmission of medical image deduplication system that can locate and eliminate the duplicate entries in massive datasets is the main goal of this project. . In addition to taking up expensive storage space, duplicate data taints data analysis, creates inconsistencies, and threatens the dependability of electronic systems. In this part, we provide an overview of earlier studies in the topic of image steganography. The method of adding information to digital content without resulting in a perceptual decline. Watermarking, stenography, and cryptography are three well-known methods of data concealment [3].

Several data hiding strategies are included in the spatial domain method for an image. When compared to the other methods, LSB substitution is the most straightforward and widely applied. Because there is just one modification applied at the LSB point, "1," the cover image is less hazy when using this method. Better PSNR and a lower MSE value are provided by the LSB approach [4]. We discovered multiple methods for using color graphics to conceal data.

We present a novel approach in this work that enables reversible data hiding in encrypted images. In this approach, a small block (B × 2 pixels) from the encrypted image can have one extra data bit embedded by the data hider. The encrypted image's non-overlapping portions, or blocks, will all be processed by accessing those blocks in a predetermined order. There is no need to change any pixel values in order to incorporate the bit value 0 in an encrypted image block. All of the pixels in the first column of the chosen image block will be mapped if bit value 1 is to be embedded.

The overall pixel distribution in the image will result from the new mapping operations we are doing on the encrypted images. The true question is whether the modifications we make to the encrypted image will have an impact on its security. The resultant encrypted images undergo both an entropy and a histogram analysis. The fact that the entropy is so close to 8 and the histogram is nearly flat suggests that the new approach won't significantly compromise the security of the picture encryption [5].

Efforts to enhance sample efficiency and maintain robustness in various machine learning and reinforcement learning algorithms have been ongoing in recent years. One approach involves the incorporation of off-policy samples, which are data points collected from policies different from the one currently being evaluated or improved upon. By leveraging off-policy samples, algorithms can potentially reuse data more effectively, leading to improved learning efficiency. Moreover, higher-order variance reduction techniques have been explored to further enhance sample efficiency and stability. These techniques aim to mitigate the variance of estimators used in learning algorithms, which can be a significant challenge, especially in high-dimensional or complex environments. The study of invisible communication, or steganography, usually focuses on how a message is hidden. Sending data across any media requires protection, which is why steganography was developed to safely transmit data in an image that is invisible to the human eye. This essay combines steganography and cryptography using an image processing method. This article proposes a YCbCr color model based on 2-bit XOR [6].

The suggested system, which transforms an image from RGB to YCbCr space and then uses 2-bit XO to bury secret data inside the Cr color space component, is an extremely safe method for data hidden in the spatial field for picture steganography.

This provide a secondary encryption secure transmission strategy to protect the original private information and guarantee quality-of-service provisioning for the secondary system. According to the suggested plan, the secondary system can gain certain spectrum opportunities while the primary system encrypts the primary confidential messages using the secure secondary messages [7]. In particular, when the primary system is secure, the primary information can be transmitted directly; if not, the spectrum will be used for secondary transmission. In the event that the primary system is insecure but the secondary messages can be transmitted securely, the primary system encrypts the primary information using the secure secondary messages [8].

III. METHODOLOGY

Proposed System: The primary goal of this research is to increase storage economy and retrieval speed in the Hadoop environment by eliminating redundant data through the use of hash-based deduplication. Hadoop is a platform for processing enormous volumes of data. It is suggested to use the DBase up approach to deduplication the stored data. Upon uploading a file to the database, the user's file is first separated using the partition algorithm file split into predetermined size chunks.

The chunks are indexed into the database in the following phase, which involves creating a hash value for each piece using the SHA 256 method. Only one unique chunk is stored if the hash value of the partitioned pieces matches the hash value of the previously stored chunk.

Benefits:

1. Businesses that use and manage their data assets well have an advantage over competitors in their particular markets. Businesses may take advantage of the full potential of their data and maintain an advantage over rivals by putting sophisticated deduplication procedures into practice [12].
2. The company can provide a better client experience when the data is accurate and consistent. Clean data increases customer happiness and loyalty in a number of ways, including more effective customer service, customized product suggestions, and targeted marketing efforts.
3. Data analytics and business intelligence projects benefit greatly from having clean, duplicated datasets as a base. Through precise and trustworthy data analysis, the company may better understand consumer behavior, market trends, and operational performance, which helps with strategic planning and well-informed decision-making.
4. The insights derived from the analysis can be translated into user-friendly applications. These applications can provide individuals with easy-to-understand feedback, making it simpler for them to make informed choices about their data.

Methodology: System Architecture:

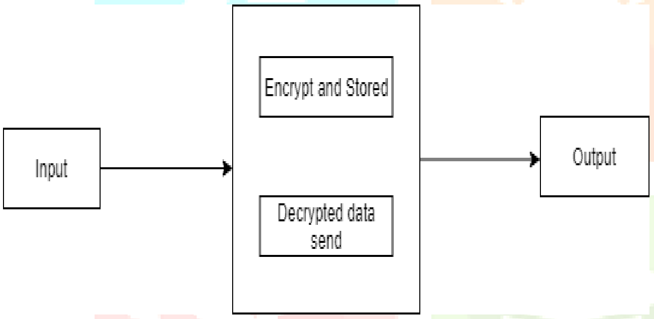

Figure 1 – System Architecture

Define Objectives and Requirements: Clearly state the project's goals, the kinds of data to be duplicated (text, photos, or both), and the results that are anticipated

1. Data Collection and Preparation: Gather unprocessed data from pertinent sources, making sure it includes a variety of text and image examples.
2. Feature Extraction: To represent text and image data quantitatively, extract features from both sources.
3. Machine Learning Model Development: Develop machine learning models for jobs involving the duplication of image and text.
4. Integration and fusion: A single combination framework by prediction of text and images duplication.
5. Model training: Train the model using the training set.
6. Model evaluation: Evaluate the model's performance on the test set

3.1 Algorithm:

Reinforcement Learning:

Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions [15]. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty. Reinforcement learning (RL) is typically used for dynamic decision-making tasks, where an agent interacts with an environment to learn optimal actions.

MD 5:

Create a byte stream for every instance of data (text document, image file, etc.) Create a distinct hash value for every data instance by applying the MD5 hashing algorithm to the byte stream. Examine the MD5 hashes produced for various data sets. There is a strong probability that the related data instances are duplicates if two hashes are identical. Data instances with similar MD5 hashes should be flagged or marked as possible duplicates. It is optional to carry out extra verification procedures or inspections to verify the duplicates before to deletion. Eliminate from the dataset any duplicate data instances with the same MD5 hash based on the comparison results.

Random Forest Regression:

What it does: Random Forest Regression (RFR) is a machine learning technique used for predicting continuous numerical values. It constructs an ensemble of decision trees and aggregates their predictions to make accurate forecasts.

How to use it: RFR can assist in choosing the best dietary and exercise options by analyzing clear choices like different diets or exercise routines. RFR aims to help individuals maintain a healthy BMI by leveraging its predictive capabilities. It learns from past experiences, adapting its predictions over time to promote healthier choices in terms of nutrition and physical activity.

Mathematical Model:

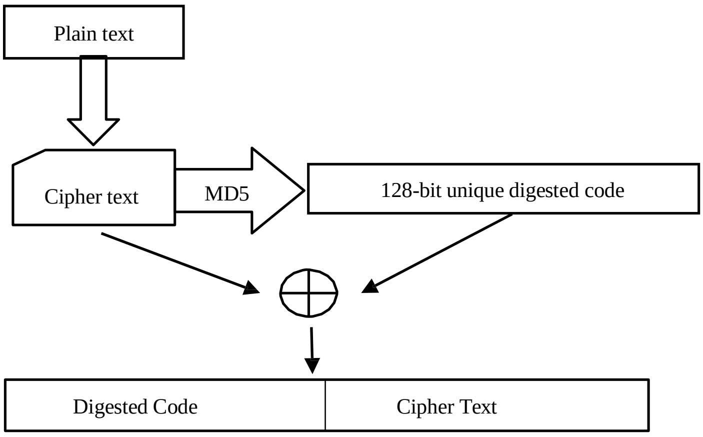

Figure 2 – Mathematical Model

IV. CONCLUSION

We covered in this paper how to use the encryption and decryption approach to prevent duplication. Additionally, we are employing three algorithms for the text uploading. The Structural Similarity AES Algorithm is what we use for uploading files to the cloud, and its primary function is to assess the quality of the image by measuring variables like brightness, contrast, and structure [13]. The two images resemblance. We use encryption to store big amounts of data efficiently and to prevent duplicate text and images. The suggested approach offers a scalable, effective, and unified method for removing duplicate data using machine learning techniques, which is a substantial improvement in the field of data management [14]. Organizations may drive innovation, realize the full value of their data assets, and achieve a competitive advantage in today's data-driven market by putting this approach into place. The solution improves organizational data dependability and integrity by eliminating duplicate records, making analyses more precise and insightful. Organizations handling big volumes of data can save money by optimizing their storage resources, freeing up funds for other important strategic projects.

V. ACKNOWLEDGEMENT

We would also like to thank Assistant Prof. Kavyashree H. N., our project guide, for mentoring us during the project work. We would also want to thank our parents and associates for their invaluable support and encouragement.

VI. REFERENCES

[1] Schütze, H., Raghavan, P., and Manning, C.D. (2008). Information Retrieval Overview. Cambridge University Press.

[2] In 2014, Leskovec, J., Rajaraman, A., and Ullman, J.D. Extensive Data Mining. Cambridge University Press.

[3] Zaki, M.J., Ganesan, K., and Hacid, M.S. (2008). finding duplicate sets of websites. Systems of Knowledge and Information, 17(1), 1–28.

[4] Wiederhold, G., Wang, J.Z., and Li, J. (2002). Semantics-Sensitive Integrated Matching for Picture Libraries: SIMPLIcity. 23(9), 947-963, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[5] T. Mishra, A. Kumar, and S. Satapathy, "Designing and Implementing Mobile Applications with Flutter," in Proc. of 2021 International Conference on Computing and Communication Systems (IC3S), 2021

[6] Schütze, H., and C.D. Manning (1999). Statistical Natural Language Processing's foundations. Press of MIT. In 2016,

[7] He, K., Zhang, X., Ren, S., and Sun, J. Image Recognition with Deep Residual Learning. IEEE Conference on Computer Vision and Pattern Recognition Proceedings, 770-778.

[8] Jordn , M. I., Ng, A.Y., and Blei, D. M. (2003). Allocation of Latent Dirichlet. 3, 993_1022, Journal of Machine Learning Reasearch.

[9] Chen, K., Corrado. G., Mikolov, T., & Dean, J. (2013)Efficient Vector Space Word Representation Estimation preprint arXiv.

[10] Ba, J., and D. P. Kingma (2014). Adam: A Stochastic Optimization Method Method. arXiv preprint 1412.6980 arXiv:1412.

[11] R. Steinberger, B. Pouliquen and J. Hagman, “Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC”, in Proceedings of the 3rd Inter. Conf. of Computational Linguistics and Intelligent Text Processing (Mexico City, Mexico, 2002, pp. 415-424.

[12] S. Brin, J. Davis and H. Garcia-Molina, “Copy Detection Mechanisms for Digital Documents”, in the ACM SIGMOD International Conference on Management of Data (San Jose, California, USA, May 22-25 1995), 1995, pp. 398-409.

[13] A. Si, H.V. Leong and R.W.H. Lau, “CHECK: A document plagiarism detection system”, in Proceedings of ACM Symposium for Applied Computing, ACM (San Jose, California, USA, February 28 - March 1 1997), 1997, pp. 70-77.

[14] M.C. Burl, M. Weber, and P. Perona, ªA Probabilistic Approach to Object Recognition Using Local Photometry and Global Geome- try,º Proc. European Conf. Computer Vision, pp. 628-641, June 1998.

[15] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom et al. ªQuery by Image and Video Content: The QBIC System,º IEEE Computer, vol. 28, no. 9, 1995.