ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 8, Issue 6, June 2019**

# An Approach for Secured and Fast transfer data Minimization Approach of Big Data using Secure Encoding: An Overview

Dabhade Jyoti Goraksh[1], Prof. Kore Kunal Sidramappa.[2]

P.G. Student, Department of Comp Engineering, Sharadchandra Pawar college of Engineering,Otur, Pune, India

Assistant Professor, Department of Comp Engineering, Sharadchandra Pawar college of Engineering, Otur, Pune, India

**ABSTRACT :** Big data delay transmission is most time consumable, over the internet. One way to increase the efficiency of data transmission is to increase the bandwidth, which might not always be possible due to resource limitations. Another way to maximize channel capacity utilization is by decreasing the bits that need to be transmitted for a given dataset. Traditionally, transmission of big genomics datasets between two geographical locations is commonly done using general-purpose protocols, such as hypertext transfer protocol (HTTP) and file transfer protocol (FTP). In this dissertation, a novel deep learning-based data minimization algorithm is presented and aims to: 1) minimize the datasets during transfer over the carrier channels; 2) protect the data from the man-in-the-middle (MITM) and other attacks by changing the binary representation (codewords) several times for the same dataset. This innovative data minimization strategy exploits the alphabet limitation of DNA sequences and modifies the binary representation (codewords) of dataset characters by usingdeep learning-based random sampling that utilizes the convolutional neural network (CNN) and Fourier transform theory. This algorithm ensures transmission of big genomics datasets with minimal bits and latency, thereby lending to a more efficient and expedient process. To evaluate this approach, extensive actual and simulated tests on various genomics datasets were conducted. Resultsindicate that the proposed data minimization algorithm is up to 99-fold faster and more secure than the current use of the HTTP data-encoding scheme and 96-fold faster than FTP on tested datasets.

## I. INTRODUCTION

The simple approach (sequential) implements integrity verification in three steps. In the first step, a file is transferred from source to destination using preferred transfer application. Once the transfer of the file is completed and it is written to the storage at destination, the second step starts during which checksum of original file at source and transferred copy at destination are computed using desired hash function such as MD5 or SHA1. In the third and final step, checksum values of original file and transferred copy are exchanged between source and destination servers to compare. If the checksum values are the same, then file transfer is marked as finished. Otherwise, transferred copy of the file will be assumed corrupted and all three steps are executed again. The main objective of running end-to-end integrity check is to detect possible data corruption by comparing the checksum of the file at the source and destination servers. On the other hand, operating systems are designed to minimize cache misses, so if a file is recently read or written, it will be kept in the memory to optimize successive accesses to the file. This causes checksum processes to be restricted to cached copy for small files even though checksum calculation is run after file transfers. That is, operating system of destination server will cache the whole file in memory as it is being streamed from network and written to disk if the file size is smaller than free memory space. So, when checksum process starts reading the file after its transfer is completed, the process will be served from memory because of which running checksum computation after file transfer does not ensure that the file will be read from disk. Similarly, files will be cached after they are read for transfer at source server. Thus, file I/O of checksum computation will be served from cache memory.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 8, Issue 6, June 2019**

End-to-end integrity verification is introduced to reduce the likelihood of accepting corrupted data. The basic implementation of end-to-end integrity verification for data transfers works as follows: Sender first reads the file from storage and sends it to destination. Once data transfer is completed, the sender reads the file again to compute checksum using a hash algorithm such as MD5 and SHA1. Receiver does the same and computes the checksum after file is fully received and written to the storage. Finally, the sender and receiver exchange the computed checksum values and compare. If checksum values of source and destination servers are the same, then the transfer is marked as successful, otherwise the file at destination server is assumed to be corrupted and transfer is restarted. If the dataset consists of multiple files, then the transfer of next file will begin only after previous file's integrity verification is completed successfully.

## II. LITERATURE SURVEY

According to a study by Forrester Research [1], 77% of the 106 large IT organizations operate three or more datacenters and run regular backup and replication services among these sites. More than half of these organizations have over a petabyte of data in their primary datacenter and expect their inter-datacenter throughput requirements to double or triple over the next couple of years.

Xiong et al. [2] proposed fsum that uses bloom filter compute the checksum of very large datasets stored in long term archival storage. Instead of calculating checksum for each file, they partition files into blocks such that multiple threads can process different portions of the file simultaneously. To achieve
ordering among threads, bloom filter is used to combine the results from threads. Once all the blocks are processed and corresponding hash values are inserted into the bloom filter, final checksum is calculated by computing the hash of the bloom filter. fsum runs up to 4x faster than traditional file-level checksum computation approach but comes with the cost of probability of false positive identification due to relying on bloom filter. On the other hand, as data transfer rates are rapidly increasing, traditional integrity verification mechanisms fall short to detect corruption. For example, some of the components of data transfers have built-in integrity verification mechanisms, however they are either weak or applicable to subset of available systems. For example, TCP checksum fails to detect errors once in 16 million to 10 billion packets [3]. As a result, researchers observed up to 5% data corruption for file transfers in 100 Gbps network [4].

Globus [5] supports end-to-end integrity verification for data transfers. It can pipeline data transfers and checksum computation to minimize the overhead of integrity verification. However, it calculates the checksum of files after their transfer completes. That is, files are read twice in the source server, one to send to destination and one to compute checksum. Yet, its pipelining approach fails to work well when dataset consist of mixed file sizes.

Liu et al. [6] proposed block-level pipelining to achieve better overlapping of checksum computation and file transfer operations. It reduces execution time considerably especially when dataset is composed of files with mixed sizes. Similar to Globus, block-level pipelining also reads files twice at source server. However, as opposed to Globus, its second read is faster for all file sizes since blocks are small enough to be kept in memory for some time, so once a block is read for transfer operation, it will be cached. When checksum computation process attempts to read the file block, it will find it in the memory. On the contrary, FIVER reads files once and run the transfer and checksum computation processes simultaneously, reducing I/O overhead and time required to compute checksum.

KissiMireku Kingsford et. al. [7] proposed a system A Mathematical Model for a Hybrid System Framework for Privacy Preservation of Patient Health Records. System presents a mathematical version for identification based encryption protocol for privacy upkeep of the affected person all through the collection of patient health facts for analysis. This has come to be an necessary part of human every day life in which health records are submitted for evaluation. The version delinks the affected person's identity from the examined records throughout records submission for the preservation of the patient's privacy.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Xueping Liang et. al. [8] proposed a system Integrating Blockchain for Data Sharing and Collaboration in Mobile Healthcare Applications, which describes an progressive user-centric health statistics sharing solution via utilizing a decentralized and permission blockchain to protect privacy the use of channel formation scheme and beautify the identity control using the membership carrier supported by the blockchain. A cell software is deployed to collect health records from private wearable devices, manual input, and scientific gadgets, and synchronize statistics to the cloud for records sharing with healthcare companies and medical insurance organizations. To hold the integrity of fitness records, inside every record, a proof of integrity and validation is permanently retrievable from cloud database and is anchored to the blockchain network. Moreover, for scalable and performance considerations, gadget undertake a tree-based facts processing and batching approach to address huge data sets of personal health data collected and uploaded by the various web platforms.

R. Manoj et. al. [9] describes a Hybrid Secure and Scalable Electronic Health Record Sharing in Hybrid Cloud, Which achieved the two green encryption techniques are combined for pleasant grained get admission to control and safety of data privacy. Multi-authority and Key-based encryption schemes are used for the encryption of every part of fitness statistics after dividing those statistics the use of a vertical partitioning approach. Multi-authority encryption schemes are ordinarily used in the Public Domains (PUDs), while Key-based encryption schemes are prevalent in Personal Domains (PSDs). Together, they provide; secure data access and authentication of users.

Shahidul Islam Khan et. al. [10] proposed a Privacy and Security Problems of National Health Data Warehouse: A Convenient Solution for Developing Countries, it gives a realistic solutions Global Patient Identification Technique (GPIT) which can anonymize identifiable private facts of the sufferers whilst maintaining document linkage in integrated health repositories to facilitate expertise discovery system. We have used encrypted cell wide variety, gender and NAMEVALUE of sufferers to supply Global Patient Identification Key. This gadget is being carried out in Bangladesh to broaden National Health Data Warehouse. Our approach is also suitable for the developing countries where poverty and illiteracy rates are high among mass people.

## III. PROPOSED SYSTEM DESIGN

The proposed data minimization mechanism relies on a deep learning-based method, while encoding the data during data transfer, and then transfers the data securely in a shortened time using Hadoop distribution file system.

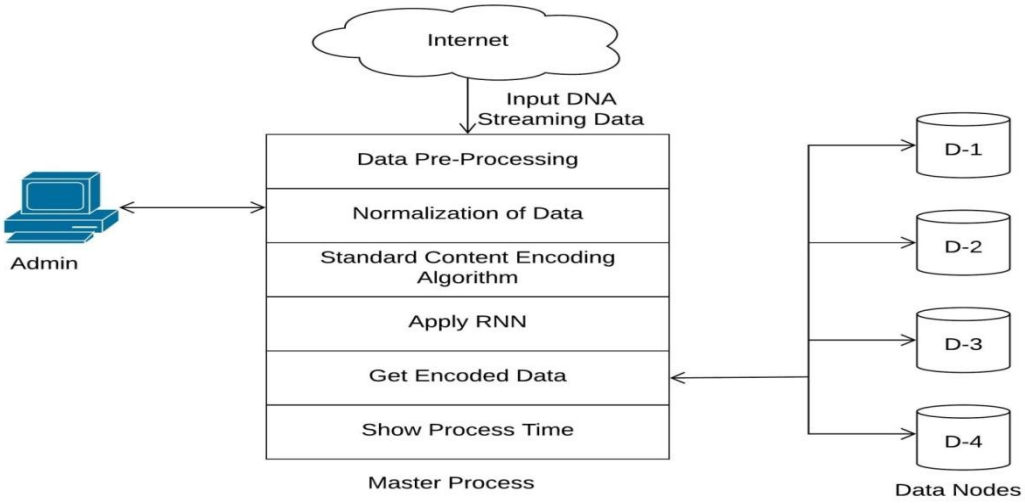


**Figure 1: Proposed System Architecture**

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 8, Issue 6, June 2019**

The proposed system develop and implement transfer protocols equipped with a novel data minimization algorithm for big genomic datasets that aim to share the data in less time and with more security. Moreover, these protocols will introduce a generic concept that can be modified to transfer securely minimum datasets that have limited symbols by using RNN -based algorithm content-encoding schemes. The system also carried out implemented a novel deep learning based data minimization algorithm to integrate with transfer protocols to reduce the size of big genomic datasets during the transfer phase, and then to transfer the data securely in less time. The implementation results illustrate that the proposed data minimization algorithm is capable of reducing the transfer time 99-fold, compared to the standard content encoding of HTTP, and 96-fold compared to FTP on tested datasets. In this paper, used new Compress algorithms as optional compression algorithms, in addition to data minimization algorithm to assess how the transfer protocol behaves in terms of transfer time and size. Also, showed that proposed data minimization algorithm provides the best size reduction, reduces transfer time, and securely transfers big genomic datasets.

## IV. CONCLUSION

It is not a minor challenge to implement a new cryptography algorithm by combining two different techniques. Also, the high demand on small, friendly, and affordable wearable health devices encourages manufacturing production. However, producing wearable devices to provide many services remotely, such as cloud-based health services, creates security challenges. Data security and privacy are very important to both users and service providers, especially for remote healthcare services. Many solutions have been developed for challenges of remote diagnostic devices. However, security and cryptography challenges have not been addressed at the same level, mainly due to compatibility issues. As a result, scientists are motivated to navigate and search for new methods to transfer health data more efficiently and in a fully secured environment. Current data encryption methods are not suitable for the cloud-based services such as remote healthcare monitoring due to using heterogeneous devices that use a variety of transfer protocols that belong to different vendors. Observing these facts, we take advantage of the nature of the genomic encryption and the deterministic Chaos Theory to implement a more efficient cryptography algorithm to secure remote healthcare monitoring.

## REFERENCES

[1] F. Research, "The Future of Data Center Wide-Area Networking," info.infineta.com/l/5622/2011-01-27/Y26
[2] S. Xiong, F. Wang, and Q. Cao, "A bloom filter based scalable data integrity check tool for large-scale dataset," in Parallel Data Storage and data Intensive Scalable Computing Systems (PDSW-DISCS), 2016 1st Joint International Workshop on. IEEE, 2016, pp. 55–60
[3] J. Stone and C. Partridge, "When the CRC and TCP checksum disagree," in ACM SIGCOMM computer communication review, vol. 30, no. 4. ACM, 2000, pp. 309–319.
[4] R. Kettimuthu, Z. Liu, D. Wheeler, I. Foster, K. Heitmann, and F. Cappello, "Transferring a Petabyte in a Day," Future Generation Computer Systems, 2018.
[5] "Globus," https://www.globus.org/.
[6] S. Liu, E.-S. Jung, R. Kettimuthu, X.-H. Sun, and M. Papka, "Towards optimizing large-scale data transfers with end-to-end integrity verification," in Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016, pp. 3002–3007.
[7] Kingsford KM, Zhang F, Ayeh MD, MaryMargaret A. A Mathematical Model for a Hybrid System Framework for Privacy Preservation of Patient Health Records. InComputer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual 2017 Jul 4 (Vol. 2, pp. 119-124). IEEE.
[8] Liang X, Zhao J, Shetty S, Liu J, Li D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. InPersonal, Indoor, and Mobile Radio Communications (PIMRC), 2017 IEEE 28th Annual International Symposium on 2017 Oct 8 (pp. 1-5). IEEE.
[9] Manoj R, Alsadoon A, Prasad PC, Costadopoulos N, Ali S. Hybrid secure and scalable electronic health record sharing in hybrid cloud. In2017 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud) 2017 Apr 6 (pp. 185-190). IEEE.
[10] Khan SI, Hoque AS. Privacy and security problems of national health data warehouse: a convenient solution for developing countries. InNetworking Systems and Security (NSysS), 2016 International Conference on 2016 Jan 7 (pp. 1-6). IEEE.