# Secured and Fast transfer and minimization of Big Data using Adaptive Encoding

DabhadeJyoti Goraksh[1] Prof. KoreKunal Sidramappa.[2]
*P.G. Student, Department of Comp Engineering SharadchandraPawar college of Engineering, Otur, Pune*
*Assistant Professor, Department of Comp Engineering, SharadchandraPawar college of Engineering, Otur, Pune*

***Abstract:*** *High dimensional data processing is the most essential task in bigdata environment there are many systems in existing offences which is already define to process the large data. The proposed work carried out to process high dimensional genomic data set in the distributed environment. Initially system works with DNA sequential data which having various attributes. In this work, designed a data minimization algorithm to transfer big genomic datasets in an expedient, secure way to allow scientists to share their data and analyses. The HTTP is used as a baseline protocol to compare and assess implementation results of transferring big genomic datasets. System algorithm provides improvements in response and transfer time of genomic datasets, as well as prevents unauthorized individuals from accessing the file contents in the event of a data breach because there assign different code words to the same character of the dataset in different times and file parts based on data obtained in running the Recurrent neural network (RNN) approach. This heuristic model, implementation results proved that proposed data minimization algorithm reduces significant amounts of data and makes more efficient use of network bandwidth, while also protecting the data by preventing unauthorized individuals from accessing them.*

***Keywords:*** *Machine Learning, Deep Learning, RNN, Mapreduce, HDFS, Hadoop***.**

## I. Introduction

Beyond the big data analytics, IoT data calls for another new class of analytics, namely fast and streaming data analytics, to support applications with high-speed data streams and requiring time-sensitive (i.e., real-time or near real-time) actions. Indeed, applications such as autonomous driving, fire prediction, driver/elderly posture (and thus consciousness and/or health condition) recognition demands for fast processing of incoming data and quick actions to achieve their target. Several researchers have proposed approaches and frameworks for fast streaming data analytics that leverage the capabilities of cloud infrastructures and services [8], [9]. Basically DNA sequencing is needed in the most critical areas such as illegal investigations, genotyping and Disease, mutation analysis, and screening of single nucleotide polymorphisms (SNP) due to disease-related genes or agents, detect chromosomal abnormalities. To identify disease- and/or drug-associated genetic variants to advance precision medicine. Also, the use of high-throughput DNA is sequencing instruments, such as next-generation sequences (NGS) technologies that include whole-genome sequencing (WGS) and whole genome sequencing (WES), significantly decreases the sequencing costs and enables the genomic datasets to join the big data club. Those instruments became big data generators, not only for bigbiology centers, but also for small biology laboratories and researchers.DNAsequences alignment system from time to times expertise complexness thanks to exponentially increase of deoxyribonucleic acid sequences knowledge. The complexness will be factorized into cluster, i.e. time and memory area however each issue has relations.

This system provides decrease the strategy exploits the alphabetic range of DNA sequences, and on the basis of deep learning, the dataset modifies the binary illustration (codew) of the characters. convolutionalneural network (CNN) to ensure a minimum of code word usesto the high frequency characters at different time slots duringthe transfer time.

## II. Litarature Survey

Weigang Li et al. [1] they analysedthe performance characterization of the compression workload in genomics analysis pipeline, and we develop A new hardware acceleration method that efficiently compresses DNA sequences with less computation time and CPU usage. The new optimization method results in a 1.5Xspeedup (1-thread) and 3.5XCPU cost savings (32-thread) for an out-of-core sorting of genomics data.

Winston Haaswijk et al. [2]they adapted the argument as a deterministic Markov decision process (MDP). Then we take advantage of recent advances in making intensive reinforcement to create a system that learns how to navigate this process. There are many desirable properties in our design. It is autonomous because it automatically learns and does not require human intervention. It generalizes for large tasks after training on small instances. In addition, it supports both internal and multilateral actions, without the need to handle special cases. Finally, this is normal because the same algorithm can be used to achieve different optimization objectives, such as size and depth.

Umberto Ferraro Petrillo et al. [3]They discuss how the development of distributed and the development of Big Data Management technologies has influenced the analysis of large datasets of biological sequences. Apart from this, we show that in relation to the specific designs with the configuration of different parameters and careful engineering of the software, the software can be important for achieving good performance, especially on very large amounts of data. We choose to count k-mers as a case study for our analysis, and with the arbitrary values of SPARK, Kashmir as a framework for implementing FastKmer, extraction of k-mer data from a large collection of organic sequences A novel approach to One of the most relevant contributions to Fastkemar is the introduction of a module to balance the aggregation of data on the nodes of the computing cluster, to eliminate data skew while allowing for the complete exploitation of the underlying distributed architecture. . We also present the results of comparative experimental analysis that our approach is currently the fastest among people based on Big Data Technologies while demonstrating very good scalability. They provide evidence that the technologies like Hadoop or Spark are used to analyze large datasets of organic sequences only when the strange aspects of Vastu details and thoughtful structures for algorithm design and implementation are kept in mind.

Haoran Sun et al. [4]they address both the theoretical and practical aspects of DNN-based algorithm approximation with wireless resource management applications. We first pin a section of the optimization algorithm, which are 'learnable' in principle by fully connected DNN. Then, we focus on the name of a popular power allocation algorithm on DNN-based approximation. They can be quite precise to confuse the DNN to approximate WMMSE - approximation error _ lightly depending on the number of layers of neurons and DNN (1 = _) in the order of logs. On the implementation side, we use a comprehensive numerical simulation to demonstrate that DNNs can obtain magnitude speed orders in computational time compared to the state-of-the-art power allocation algorithm based on optimization.

RuchieBhardwaj et al. [5]the author works on emerging large data and analytics with opportunities, the progress and challenges of genomics. As the productivity of technology development and healthcare industry increases, the number of people benefiting from the healthcare industry is increasing. Five Way Pathways: Defining the right life, the right care, the right provider, the right value and the right innovation, the structure of the new industry. It has been proved that by adopting this new approach to health care, only $ 1 billion can be saved in one health facility and up to 450 billion dollars in the United States can be saved. In addition, the introduction of equipment which collects large amounts of data, leads to evidence-based and preventive-based medicines, which is going through the industry. These approaches lead to a more successful treatment for patients.

Tariq Abdullah et al. [6]they presented a genomics data analysis framework that addresses each the problems of existing genetic science analysis pipelines. It reads unstructured genetic science information from sources, transforms it in a very structured format and stores this information into a No-SQL info. during this method, genetic science information is queried associatey|likeall|likeseveral|like every} different information and an update within the genetic science information doesn't need reading the entire information set. The framework conjointly presents Associate in Nursingeconomic analysis pipeline for analyzing the genetic science information for a range of functions like genotype bunch, organic phenomenon microarrays, body variations or inheritance analysis. A case study of genotype bunch is conferred to demonstrate and assess the effectiveness of the conferred framework. Our results show that the framework improves overall performance of the genetic science information analysis pip line by forty ninth from existing genetic science information analysis pipelines. moreover, our approach is powerful and is ready sustain high performance with high system workloads.

FabrizioCelesti et al. [7] they analyze and classify the major current deep learning solutions that allow biotechnology researchers to perform big genomics data analytics. Moreover, by means of a taxonomic analysis, we provide a clear picture of the current state of the art also discussing future challenges.

They presented taxonomy in order to analyze the current state of the art of NGS software solutions based on Deep Learning algorithms. From our study, we realized that Deep Learning solutions for NGS are at an early stage. Many researchers are beginning to look at Deep Learning especially in context of regulatory sequences and DNA binding proteins,but there are not yet so many solutions focusing on other genomics research fields. Moreover, from our analysis, it is evident how Deep Learning is currently adopted in NGS standalone applications. In our opinion, the research community could benefit of the advantages offered by future NGS Cloud computing services based on Deep Learning algorithms in terms of both computational resource scalability and DNA fragments data sharing.With this paper, system succeeded in stimulating the biotechnology community toward the development of new emerging NGS Cloud platform and software based on Deep Learning.

Mohammed Aledhari et al.[8] They presents a new real-time data minimization mechanism of big genomic datasets to shorten the transfer time in a more secure manner, despite the potential occurrence of a data breach. Our method involves the application of the random sampling of Fourier transform theory to the real time generated big genomic datasets of both formats: FASTA and FASTQ and assigns the lowest possible code word to the most frequent characters of the datasets. Our results indicate that theproposed data minimization algorithm is up to 79% of FASTA datasets size reduction, with 98-fold quicker and further secure than the standard data-encoding method. Also, the results show up to 45% of FASTQ datasets size reduction with 57-fold faster than the standard data-encoding approach. Based on our results, we conclude that the proposed data minimization algorithm provides the best performance among current data-encoding approaches for big real-time generated genomic datasets.

Marco Masseroli et al. [9] They proposed a exampleshift in genomic knowledge management, supported the Genomic Data Model (GDM) for mediating existing knowledge formats and on the Geno Metric source language (GMQL) for supporting, at a high level of abstraction, knowledge extraction and therefore the commonest knowledge-driven computations needed by tertiary data analysis of Next Generation Sequencing datasets. Here, we tend to gift a brand new GMQL-based system with increased accessibility, immovableness, quantifiability and performance. this can be the new system includes a well-designed standard design featuring: i) associate degree intermediate illustration supporting many various implement ii) a high-level technology-independent mine abstraction, backup completely different repository technologies (e.g., native filing system, Hadoop filing system, database, or others); iii) many system interfaces, as well as an easy internet primarily based interface, an internet A program interface for ser ice interface, and Python language. Examples of biological use cases, afflicted public ENCODE, roadmap epnomomics and TCGA datasets show the relevancy of their work

Karen Y. He et al. [10] Review the challenges of manipulation of next-generation information (NGS) information and many clinical information on the next generation of generic drugs generated from EHR for generic drugs. We have a tendency to introduce potential solutions for various experiments in deploying, supervision, and evaluating genomic and clinical information to implement genomic drugs. To boot, we have a tendency to conjointly gift a sensible massive information toolset for characteristic clinically unjust genetic variants victimization high-throughput NGS information and EHRs. EHRs square measure exceptionally non-public, strategies of protective patient information ought to ensure that patient data is just shared with those with licensed access. Even with the present challenges, the potential blessings that genomic information will rouse attention square measure way more vital than the potential disadvantages. The increasing development of integration genomic information with EHRs might cause issues, however genomic information will definitely play a very vital role in advancing genomic drugs solely if patient privacy and data security can be exactingly protected.

Wesam H. et. Al. [11] describes the system PHeDHA: Protecting Healthcare Data in Health Information Exchanges with Active Data Bundles, which proposes a HIE system called PHeDHA (Protecting Healthcare Data in HIEs with Active Data Bundles), which provides privacy and security protection for patient data during their transmission via an HIE among different healthcare providers. PHeDHA uses as its basis the scheme named Active Data Bundles with Trusted Third Party (ADB-TTP). As the name suggests, ADBTTP is based on an integration of a trusted third party (TTP) with Active Data Bundles (ADBs). This system also presents the basic structure and operations of the PHeDHA system for protecting patient healthcare data exchanged among different healthcare providers in a health information exchange (HIE). PHeDHA aims to alleviate privacy and security concerns related to these data exchanges.

KissiMireku Kingsford et. al. [12] 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.

Xueping Liang et. al. [13] 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. Manojet. al. [14] 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. [15] 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 Methodology

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.

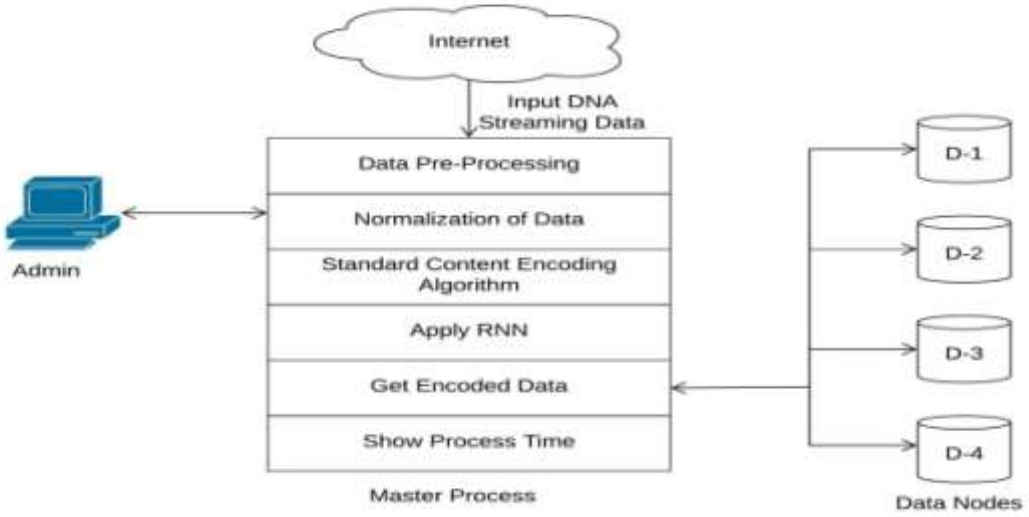

**Figure 1: Proposed System Architecture**

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 totransfer 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 baseddata 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. System Analysis

**Algorithms used**

**Algorithm 1: Dynamic Variable-length binary encoding**

**Input: Plain data with multiple attributes A[i……n], Encoding policies.**

**Output: Encoded Optimized String EC[]**

**Step 1 :**Read data from current file system using below formula

$$\text{Data} = \sum_{k=0}^{n} \{\text{DataAttributes}[i] \ldots . [n]\}$$

**Step 2 :**Read all encoding policies

**Step 3 :** Encode data using below formula

EncData[] = {Data ← policy{CharAt[0…n]}}

**Step 4 :**compressEncData[] and send into network stream

**Step 5 :** Apply decryption on EncData[].

DecData= {EncData←policy{CharAt[0…n]}}

**Step 6:**RturnDecData

**System Requirements**

1. System interfaces: Ubentu Operating System
2. User interfaces: JSP, servlet java environment

**3. Hardware interfaces:**

Processor :- Intel R-Core i3 2.7 or above

Memory :- 4GB or above

Hard Disk :- 500 GB

**4. Software interfaces**:

Front End: JSP, servlet,IE 7.0/above

Back-End: MYSQL 5.1, Hadoop 1.0

Server : Apache Tomcat

**5. Communications interfaces**

System will use HTTP as well as SMTP and SOAP protocol for establishing connection and transmitting data over the network.

**6. Services**: Amazon EC2 as Public cloud Environment (optional)

## V. Results And Discussion

The below figure shows Transfer time in millisecond of 500 MB genomic dataset using the proposed data minimization algorithm using a single machine and a standard content-encoding algorithm using multiple 1 to 10 machines in parallel.

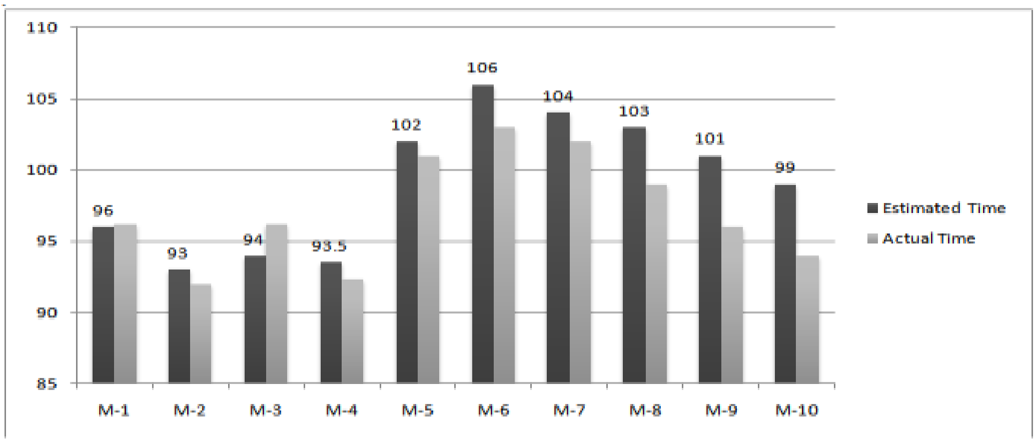

**Figure 2 :Time required (in Milliseconds) for transfer the genomic dataset using proposed algorithm from machine 1 to 10**

According above experiment illustrated in figure 2, illustrates the proposed data reduction and it's time for data transmission, it also shows the time required for estimated as well as actual time taking by system.

## VI. Conclusion

In this work implemented a novel deep learning based RNN data minimization algorithm to integrate with transfer protocols to reduce the size of big genomic datasets during the transfer phase in HDFS environment. The mapreduce first process whole data, and then to transfer the data securely in less time.

Proposed Deep learning-based neural network algorithm can process data in distributed environment. For the proposed research work we done various experiment analysis first we created 10 machines setup to measure the system performance. 500 MB data has device into 10 fold, 15 fold and 20 fold respectively. Before executing the algorithm we calculated some estimated time for transmission of data into the network stream. One data has received to destination machine we evaluate the actual arrival time with estimated time. According to the whole experiment analysis system illustrates how proposed algorithm provides higher accuracy and minimum time complexity to the system.

## References

[1]. Li W. Optimize Genomics Data Compression with Hardware Accelerator. InData Compression Conference (DCC), 2017 Apr 4 (pp. 446-446). IEEE.

[2]. Haaswijk W, Collins E, Seguin B, Soeken M, Kaplan F, Süsstrunk S, De Micheli G. Deep Learning for Logic Optimization Algorithms. InCircuits and Systems (ISCAS), 2018 IEEE International Symposium on 2018 May 27 (pp. 1-4). IEEE.

[3]. Petrillo UF, Sorella M, Cattaneo G, Giancarlo R, Rombo S. Analyzing Big Datasets of Genomic Sequences: Fast and Scalable Collection of k-mer Statistics. arXiv preprint arXiv:1807.01566. 2018 Jul 4.

[4]. Sun H, Chen X, Shi Q, Hong M, Fu X, Sidiropoulos ND. Learning to optimize: Training deep neural networks for wireless resource management. InSignal Processing Advances in Wireless Communications (SPAWC), 2017 IEEE 18th International Workshop on 2017 Jul 3 (pp. 1-6). IEEE.

[5]. Bhardwaj R, Sethi A, Nambiar R. Big data in genomics: An overview. InBig Data (Big Data), 2014 IEEE International Conference on 2014 Oct 27 (pp. 45-49). IEEE.

[6]. Abdullah T, Ahmet A. Genomics Analyser: A Big Data Framework for Analysing Genomics Data. InProceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies 2017 Dec 5 (pp. 189-197).ACM.

[7]. Celesti F, Celesti A, Carnevale L, Galletta A, Campo S, Romano A, Bramanti P, Villari M. Big data analytics in genomics: The point on Deep Learning solutions. InComputers and Communications (ISCC), 2017 IEEE Symposium on 2017 Jul 3 (pp. 306-309). IEEE.

[8]. Aledhari M, Di Pierro M, Saeed F. A Fourier-Based Data Minimization Algorithm for Fast and Secure Transfer of Big Genomic Datasets. In2018 IEEE International Congress on Big Data (BigData Congress) 2018 Jul 1 (pp. 128-134). IEEE.

[9]. Masseroli M, Canakoglu A, Pinoli P, Kaitoua A, Gulino A, Horlova O, Nanni L, Bernasconi A, Perna S, Stamoulakatou E, Ceri S. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data.

[10]. [10] He K, Ge D, He M. Big data analytics for genomic medicine. International journal of molecular sciences. 2017 Feb 15;18(2):412.

[11]. Fadheel W, Salih R, Lilien L. PHeDHA: Protecting Healthcare Data in Health Information Exchanges with Active Data Bundles. In2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) 2018 Aug 1 (pp. 1187-1195). IEEE.

[12]. 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.

[13]. 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.

[14]. 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.

[15]. 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.