A ROBUST AND SECURED ENCRYPTION SCHEME FOR BIG DATA SECURITY BASED ON HADOOP DISTRIBUTED FILE SYSTEM

Chandra Shekhar[1], Dr. Manish Varshney[2]

Article History: Received: 20.05.2023 **Revised:** 06.07.2023 **Accepted:** 22.08.2023

Abstract

The last decade is of data generation, online data availability and rapid usage of social media which leads to generation of big data consequent of that cyber assaults happened between years 2017 and 2022, ranging from ransom ware to crypto currency theft, data loss to supply chain attacks. According to the Identity Theft Research Center (ITRC), data breaches increased by 17% in 2021 compared to 2020. One of the most destructive contemporary breaches was an attack on Microsoft Exchange that resulted in multiple zero-day vulnerabilities and it was found the Proxy Logon issues in January which were rectified in March, originally exploited by Hafnium hacking organization. Harvard Business Review reported a 300% rise in number of cyber-attacks and the amount of money paid by firms in 2018. Hacker groups taking advantage of the remote work on web has weakened security measures.

To ensure the communication more secure, cryptography is used to detect malicious third parties which are called as adversaries. The technique of encryption practices an algorithm along with a key (public or private) to translate a plaintext intake into a scrambled output (i.e. cipher text). The same plaintext is turned into the same decryption since the same key is used. The technique is considered safe if an adversary cannot decrypt the cipher text to determine the key or plaintext. Using numerous combinations of plaintext and cipher text, an attacker should be impossible to learn anything about a key The proposed MEAB has achieved a 0.58% reduction in encryption time compared to AES for text files, a 0.19% reduction compared to Blowfish for text files, a 3% reduction compared to AES for pictures, a 2.29% reduction compared to Blowfish for pictures, a 1.9% reduction compared to AES for pdfs, a 1.66% reduction compared to Blowfish for pdfs, and more. 6.10% less time is spent encrypting videos than Blowfish, and 6.7% less time is spent encrypting videos than AES. 3.94% faster encryption for text files as compared to AES, for word files, Blowfish encryption time was faster by 2.9%. The entropy and Hardening index is enhanced by 1.53% and 13% against Blowfish algorithm.

Keywords: Cryptography, MEAB, Entropy, Big Data, HDFS

[1]Phd Scholar, Department of Computer Science Engineering, Maharishi University of Information Technology, IIM Road, Lucknow (U.P), India

[2]Professor, Department of Computer Science Engineering, Maharishi University of Information Technology, IIM Road, Lucknow (U.P), India

DOI: 10.31838/ecb/2023.12.s3.839

1. INTRODUCTION

In today's world, data is exploding in size due to the usage of a large variety of smart devices. About 500 Terabyte data is generated through social media. The Jet engine generates millions of petabytes in fractions of time. These are a few examples that show that Big Data is an important field and have an enormous scope, as it is difficult to store, analyse and extract insights from it. These issues are the source of different research problems i.e. storage management issues, scalable privacy presentation and security in Big Data[1].
In this section, the systematic process of research is done by the researcher and is explained thoroughly. In the first place Big Data, its significance and Its research scope are well explained followed by a good description of security in Big Data and then transition to the research area Use of -State of- art Encryption Methods. The researcher justifies the Use of- State -of -art Encryption Methods by describing the issues and techniques involved in it. The whole chapter is divided into different sections and each section converses the research area and reaches a conclusion. Each section describes the research scope, applications and techniques involved in that field.
The current era is the information era and it is the end product of expeditious development in information technology now it has become the backbone of the economy. The time immortal philosophers believed that the strength that comes from interpreting information is known as knowledge. Information, in turn, is defined as meaning assigned to data[2].

Data: Generally, data is referred to as 'raw' data. It means that data is a collection of text, number, and symbol which does not convey meaning. To gain relevance and meaning of data; it is to be processed [3].

Information: When meaning is associated with Data is called information. Processed data in a specific context provides information. Processed data convert into information making it interpretable and significant[4].

Knowledge: Knowledge is accessed by applying techniques to solve problems; termed a deterministic approach which synthesizes the data and extracts useful information[5]. Gamble et al. stated that Knowledge provides the context and structure for assessing and digesting new experiences and information, which is a fluid amalgamation of ideals, background knowledge, context, professional acumen, and firmly rooted intuition.

Wisdom: Wisdom has non-probabilistic and non-deterministic nature. Once Knowledge is assimilated in the due course of time and with experience, it helps to attain wisdom in a given context. It is essential to provide functional assessments and judgments[6].
As technology has profiled its feet in every field leads to an exponential growth of data and is termed digital data. This data is generated at a colossal rate from different kinds and immeasurable sources and is having intricate structures which are not limited to handle for conventional tools and techniques and are generated as a result of various activities performed on the network rather than directly e.g. mobile technology advancement triggers the growth of "Digital Exhaust Data" at a rapid pace[7]. Due to its scope, some big organizations establish terminals for everyone to access and inspect within it to retrieve useful information and knowledge (which are explained in the above section) e.g., Google delivers freely possibility of Google Insights, although others hesitate to offer any access. Due to its growth, **Clive Humby** termed as **"Data is the new oil."**
As this Digital Data is termed Big Data and initially it is characterized with the help of

3v's i.e., Volume, Velocity and Variety. Here volume is specified as a gigantic dataset; Velocity is specified as the pace at which data is generated; As it is generated from a huge amount of sources variety indicates the diversity of data.

Presently, there are lots of definitions of big data. The most familiar version comes from IBM which recommended that big data could be illustrated by three "V" words to examine artifacts the situations, events and so on: volume, variety and velocity[8]. The origin of the term 'Big Data' is since they are establishing a massive amount of data every day. Earlier, Three variables characterize big data i.e. Velocity, Volume and Variety [9]. But with the progression of technology, big data has gained new dimensions such as truthfulness, validity, volatility and value which are more semantically applicable to large data.

Furthermore, several features connected to large data technical issues such as Variability and Visualization have evolved. The privacy and security problems that characterize today's big data have been linked to the Valence and Vulnerability dimensions in particular. These V's are explained below[8].

RELATED WORK

Thakur and Kumar [10] compared the performance of different encryption algorithms i.e. Data encryption standard, Advanced encryption standard, and Blowfish using several criteria such as block size, key size and speed. The simulation was used to implement the algorithms in different block cipher modes i.e. electronic code block, cipher block chaining, cipher feedback, output feedback and counter mode. The researcher did a comparative analysis of these modes and found out that the algorithm performs better in OFB cipher mode than in CBC cipher mode while the results come out better in the case of CBC when compared with CFB cipher mode.

Khudhair Salah et al. [21] compared the AES algorithm to the triple-DES method when used in conjunction with the blowfish algorithm. Both DES and triple-DES algorithms are vulnerable to attacks. To overcome the drawback of DES, this algorithm is combined with the advantages of triple DES and blowfish. It results in the improvement of the existing algorithm in terms of security and speed. This enhancement of the existing algorithm is considered a proposed algorithm with enhanced speed and security.

Mahmud et al. [11] proposed a hybrid method combining strengthened AES and blowfish data secrecy with MD5 data integrity. Performance is measured in terms of time, throughput, and memory usage. The Blowfish method has a very high throughput when compared to the AES and combined AES and BF algorithms.

Princy [22] gave a detailed analysis of different existing encryption algorithms. He compared various symmetric encryption methods like DES, AES, 3DES, AES, RC4 (shared key stream cipher algorithm), RC6 (symmetric block cryptographic algorithm), and blowfish algorithm. The blowfish algorithm is studied in detail and found out that. If the blowfish's key size is increased to 448, it will be able to provide better data security even when communicating via insecure channels.

Kapil et al. [25] described Hadoop Distributed File System (HDFS) storage as providing flexible and low-cost services to enormous amounts of data. The researcher also explained the lack of any intrinsic security measure in Hadoop leads to an increase in harmful assaults on the data processed or stored through Hadoop. To enhance the data security in HDFS, attribute Based Honey Encryption is a technique abbreviated as ABHE proposed by the authors. In this proposed work, on Hadoop, ABHE is coupled with honey encryption. This method works with files that have been encoded in HDFS and decoded in the Mapper. When compared to

existing methods such as AES and AES with OTP, the ABHE algorithm has exhibited exceptional performance when conducting encryption and decryption on various file sizes.

Al Neaimi & Hassan [26] outlined every process useful in improving the blowfish algorithm's performance. A framework of 16 rounds has been constructed by using substitution in the proposed technique. The most economical and convincing technique to generate alternative keys is to employ cell automata (CA). The suggested structure provides extraordinary encoding while also providing outstanding protection against cryptography-related activities.

Manikandan et al. [27] offered a software solution that significantly improves security by using a repeating technique based on the needs of the sender. When compared to a non-iterative technique, the author offered a repeated approach based on tests that improve security.

Kumar & Kathikeyan [23] worked on the two symmetric algorithms i.e. blowfish and Rejindael. The author discussed the implementation of these algorithms. Also, on behalf of the designs of these algorithms. Implementation is done and results are compared. A comparative analysis is also done in terms of performance via performing experiments.

Padmavathi & Kumari et al. [24] reported that due to the distribution of data on a network (called distributed network), data is very unsafe. So, several strategies are required to be proposed. A strategy of integrating encryption and steganography to save the data has been proposed. The author carried out 3 encoding strategies like DES, AES, and RSA. The proposed work also includes a comparative analysis of these encoding strategies.

Sandhu et al. [28] observed that traditional strategies like cryptography and steganography are getting advanced with time. Although, several pitfalls are there in these techniques. For higher security, big key size, complicated encoding approaches as well as additional computational upstairs for central processing unit has to be proposed.

Marwaha et al. [29] explained that cryptography was used to highlight the need for data protection. Encryption is useful for ensuring the privacy, verification, veracity, accessibility and identity of individual statistics as well as the security and privacy of statistics. The author looked at three different algorithms: DES, 3-DES and RSA. DES and 3-DES are symmetric key encryption algorithms, whereas RSA is an uneven key encryption algorithm. Based on cozy statistics, time spent encoding information and procedure call output, A comparison analysis was performed.

Juremi et al. [30] proposed a novel AES-like layout for key- established AES using s-container rotation. The key extension method is combined with s-field alternation in this technique. It leads to the establishment of a key-based S-box, which protects the block cipher effectively. This unique design investigated the use of the NIST arithmetical check and may also be crypt analyzed using arithmetical assault to allow destabilization or evasion.

Kirubanandasarathy et al. [31] Cryptography was explained in terms of mathematical techniques for maintaining confidentiality, statistics integrity, entity authentication and data records foundation authentication. The symmetric block cipher encrypted by AES is the most important aspect of statistic encoding. AES has the advantage of being able to be used in both hardware and software. The AES hardware implementation has several benefits, including increased throughput and increased security. The Verilog hardware descriptive language was used to create varying degrees of pipelining for 128-bit AES encryption and decryption in this proposed study. The author proposed using the AES algorithm's pipelined structure to boost overall throughput.

Patil & Goudar [32] observed that encryption translates the data from the ordinary shape into a coded form. Two important traits i.e., capacity to save the statistics from malicious attacks as well as the speed and performance of the algorithm. Different encryption procedures i.e. DES, AES and blowfish are compared on the observation of different parameters. Quickness, block size and key magnitude are the three parameters.

Sumitra [33] expounded that network security has become a top priority in recent years. Encryption has been relegated to unstructured data and controls access through the distribution of a private cryptographic key across certain devices. Encryption transforms incomprehensible communication into unknown data through a series of statistical adjustments. Encryption is the scrawling or incomprehensibility of the content of recordings such as text, photos, aural and audiovisual at some point during transmission. Its major purpose is to prevent unwanted access to records.

Acharya et al. [18] proposed that cryptography is seen to play a key role in preserving statistics, particularly in communities. The most important aspect of data security is community protection, which refers to all components and programs, features, capabilities, active approaches, duty, entrance management, managerial and administration policy. It requires encryption because it ensures confidentiality, secrecy and identification.

Ismil et al. [34] proposed an algorithm called Rijndael algorithm as an advancement over AES. The effect of rearranging the structure of AES, specifically changing constant rotation with variable rotation, is investigated in this article. Dynamic rotation is the name given to the resulting dual cipher for better encryption of dates. Dares with variable rotation enhance both the complexity of the set of rules and the time required for brute-force attacks. A comparison of the diffusion of AES and daring algorithms has been made. Dares appears to have achieved an applicable stage of spread sooner than AES.

Goyal [18] observed and focused on large data problems as well as various security issues in Hadoop's construction primary layer called Hadoop distributed file system (HDFS). The author narrated three strategies i.e. Kerberos, Algorithm and Name hub used to improve the safety of HDFS.

Parmar et al. [35] researched huge documents and how to handle them. Due to the differentiation and not gradually critical components, enormous information security difficulties arise. The author looked at all of the major and little issues with Large Documents.

It is finally proved that the security of databases is essential. To provide robust security to the database system, several security technologies as well as techniques are proposed. Following are some potential techniques and approaches with advantages as well as limitations:

Table 1: Different studies on Hadoop

Author Name	**Approach/technique**	**Advantages**	**Limitations**
Hadeer Mahmoud[36]	Performed encryption using AES & OTP integrated on Hadoop	Encrypted HDFS files result in enhancing the performance of encryption /decryptionfile	Planning to achieve the parallel encryption processes to improve the performance ofdata encryption and secure Hadoop using structure data

Shangguang Wang [38]	Integration of the HBase table and a non-functional property index mechanism with HBase and Map Reduce	Enhancement in management of web services	Planning to deal with the dependability of The service system work and deploying in the Big Cloud system
Raj R. Parmar [39]	Integration of Kerberos and SASL with Hadoop	Due to the flexibility in the kuber framework, usage of any encryption method is possible	Implementation will be on powerful machines as well as work will be done to increase the encryption speed
Girish Prasad Patro [27]	Asymmetric key cryptosystem RSA HDFS files	Robust encryption security and compression within the Hadoop framework	Trying to improve the data import mechanism as well as the infrastructure of HDFS for better processing & management of big data
Vijaykumar N. Patil [28]	Integration of AES & O Auth with HDFS files	More secured data in HDFS due to the authentication via a unique authorization token	Try to find more security securing data & job Execution
Gayatri Kapil [20]	Integration of attribute-based encryption with honey Encryption (ABHE) on Hadoop	Considerable improvement in performance in terms of reducing the size of encryption- decryption files	To find a more robust Approach for visualizing and designing

2. PROPOSED METHODOLOGY

In this part, the proposed model—the Modified Effective Approach of Blowfish Algorithm will be explained (MEAB Algorithm). The suggested technique will cope with a range of input data, including text files, videos, and PDF format, and encrypt input data that is 128 bits long rather than 64 bits.

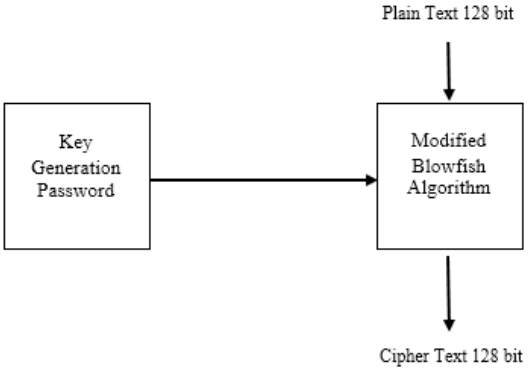

Fig 1 MAEB Algorithm with Password

The improvement keeps all the mathematical computations while producing better outcomes. Changes made to the algorithm's security, key size, and block size result in the production of strong cryptographic keys based on user-accessible passwords. By utilising the GTk hash map, MEAB algorithm also works for improved integrity, ensuring that there won't be any data modification. GtkHash is a GUI application that may be used to produce and validate checksums for a variety of methods. Hash algorithms like SHA512, MD5, SHA256, SHA1, Tiger, RipeMD, HAVAL, and Whirlpool are all supported. Launch GtkHash when it has been installed, then go to the file you want to add and select Browse from the File menu.

Put the original hash value into the Check section and click the Hash button to compare the checksum of the downloaded file to the original. The following image displays the SHA1 hash verification for the Ubuntu MATE ISO package.

Instead than storing user-provided passwords, one possibility is to generate keys from them. Due to user-supplied passwords that are not used, key creation of some kind is required. During key creation, the password is stored as a document in UTF-16 format. This file may then be referred to by Blowfish Crypt. The main goal of this technique is to make user data unavailable to everyone else and only accessible to the receipient,

Every map reduce works which transform the records must be logged. Information of users, which are in control for those jobs, must be recorded.

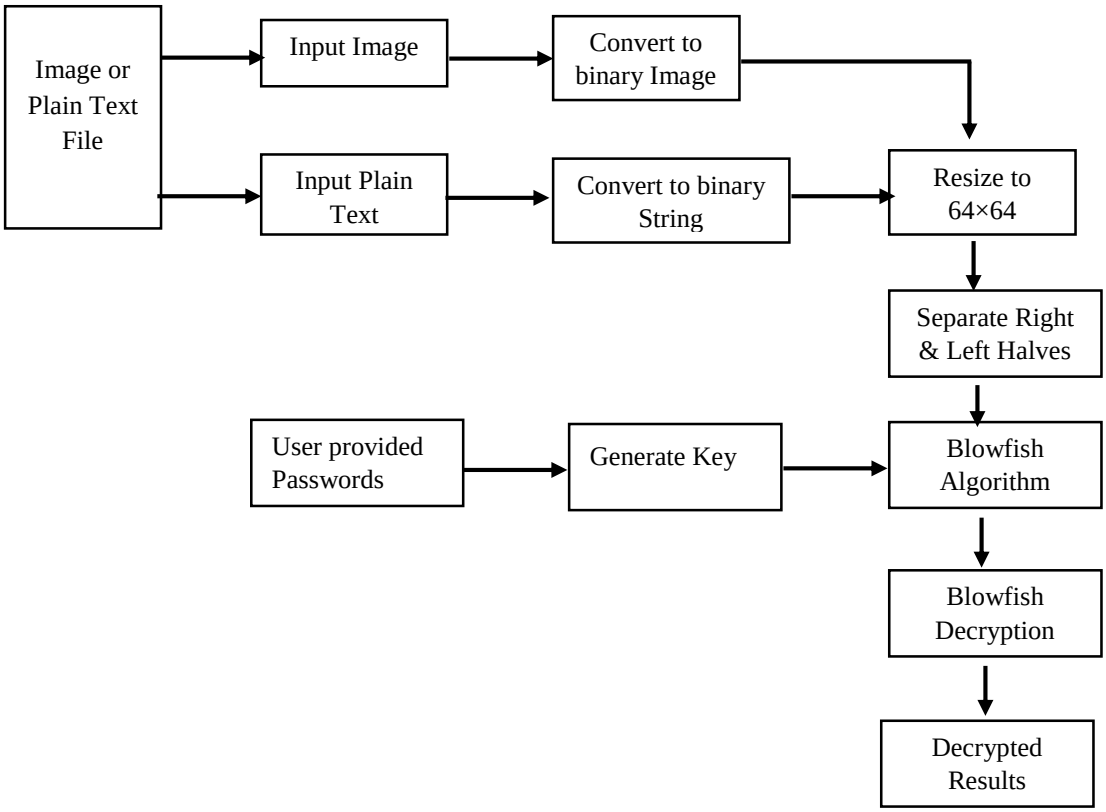

Figure 2: Flowchart of MEAB Algorithm

The secret key is XORed with the P-entries one at a time to encrypt the full zero string. The generated ciphertext is used to replace P1 and P2 by encrypting the new P1 and P2

with the updated subkeys. They are the new outputs, P3 and P4. For the P-array and S-boxes, the A Blowfish algorithm will repeat itself a total of 521 times. Around 4 KB worth of data is processed overall.

Proposed Framework

- A Proposed framework termed as Modified Effective Approach for Blowfish Algorithm (MEAB).
- Presented Framework integrate the improved Blowfish algorithm for security.
- Enhancement retains all the mathematical calculations with improved results.
- Analysis on security,size of Key, time complexity and space complexity.
- Generation of effective approach for Blowfish algorithm based on Human Achievable Passwords.
- In Symmetric Encryption, Passwords swings can 't be used as keys, so specific type of key derivation is required.

Proposed Algorithmic Approach - Modified Effective Approach of Blowfish (MEAB Algorithm)

STEP 1. S-BOX CREATION AND PRE-PROCESSING

a. All the sub-keys are calculated by using the MEAB, Size of S-box is reduced to two S-boxes

STEP 2. SPLITTING AND SEGMENTATION

a. Main difference is in the size of input block.
b. The input blocks is of128-bit and splited into 64-bit two equal segments i.e LE0, RE0.

STEP 3. CLASSIFICATION

a. The first segment LE0 is XORed with the first entry in the P-array (P 1, P11) with two 32-bit entries.

STEP 4. XOR OPERATION

a. The output is generated from Step 3 is pass to the function block i.e F..
b. Then output is generated from is combined with the second segment (RE0) of the plaintext.

STEP 5. SWAPPING

a. Swap both LE0 and RE0.
b. This process will be continue up to the eighth round.
c. After the completion of eighth round after then exchange bothLE8 and RE8 reversing the last swap.

STEP 6. RECOMBINE AND EVALUATION

a. Then RE8 is XORed to P-array (P9, P19)
b. LE8 is XORed to P-array (P10, P20)
c. Finally, the last step is recombine LE9 and RE9 to get the cipher text

3. RESULT ANALYSIS

This research work is provided with techniques based on the Blowfish encryption algorithm to address the aforementioned issue. The study focuses on providing an authentication mechanism for person validation utilising user passwords. The results of the MEAB algorithm's implementation in HDFS are shown in Section III, along with an assessment of MEAB based on the hardening index, entropy, and execution evaluation time. HDFS integrity verification is shown in Section III.
Authors have analyzed, MEAB algorithm on Hadoop with YARN Scheduler and Hadoop with Fair ScheJduler. Following table depicts the description of both schedulers.

Table 2: Selected Schedulers

Description	Associated Scheduler
Minimum delay and maximum degreeof performance	Capacity Analysis YARN Scheduler
Assigning resources in a manner thatensures that all users get an equal number of resources over time is known as fair scheduling. In Hadoop2.x, several resource types may be scheduled.	Fair Scheduler

As shown in Table 4.1 two schedulers Capacity Analysis YARN and Fair schedulers are selected for evaluation purpose.

Yarn Capacity Scheduler:
Authors have considered Average map Execution Time(sec), Total task Execution Time(sec), Virtual Memory (MB), CPU Time(sec)and Physical memory (MB) as evaluation parameter. All evaluation parameters are evaluated in Yarn Capacity Scheduler against Real time data i.e. Image file, Pdfs file, Text file, Video file and Word processing file. All the data is in MB and evaluation is shown in Table 4.

Table 4: Hadoop with Yarn Capacity Scheduler

Hadoop Applications	Average map Execution Time in Seconds	Total Task Execution Timein Seconds	Average Reduce Execution Time in Seconds	Virtual memory in MB	CPU Time (sec)	Physical memory In MB
Image Files (15)	9	23	1	695.95	10.34	445.05
Pdfs (10)	8	22	1	650.75	4.83	355.50
Text File (15)	16	39	1	675.75	4.45	252.65
Video File (10)	325	7460	1920	35875.90	1050.5	15565.90
Word Processing Files (15)	9	22	1	655.85	4.95	365.55

Average map execution time (sec), total task execution time (sec), virtual memory (MB), CPU time (sec), and physical memory (MB) were all taken into consideration by the authors as assessment criteria. In Fair Scheduler, all assessment criteria are assessed against real-time data, including image, pdf, text, video, and word processing files. The evaluation is displayed in Table 5 and all the data is in MB.

Table 5: Hadoop with Fair Scheduler

Hadoop Applications	Average map Execution Time in Seconds	Total Task Execution Time in Seconds	Physical memory In MB	Average reduce Execution Time in Seconds	CPU Execution Time in Seconds	Virtual memory In MB
Image Files(15)	10	25	335.25	1	8.35	672.75
Pdfs (10)	9	25	343.70	1	3.87	630.75
Text File (15)	9	25	255.25	1	4.1	640.70
Video File (10)	275	6558	13750.86	1668	905.79	29915.60
Word ProcessingFiles (15)	9	25	320.45	1	3.95	620.50

As shown in tables 4.2 and 4.3 Hadoop application containing variety of data such as Text, Images, Pdfs, Videos and Word Files are scheduled using Capacity Analysis YARN scheduler and Fair scheduler. Scheduler performance evaluation is done on the basis of various parameters.

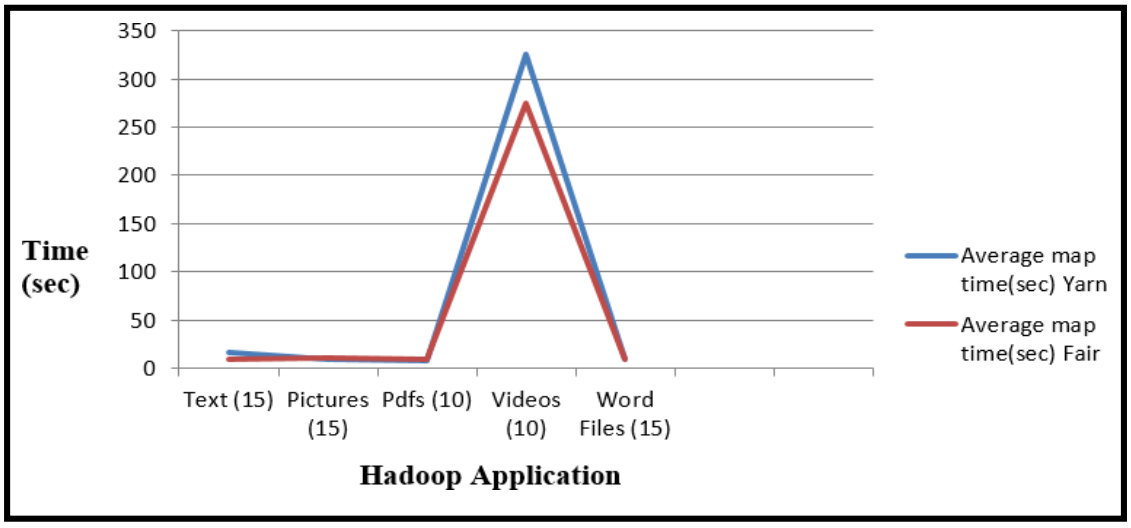

Figure 3: Capacity Analysis YARN vs. Fair Scheduler Total Task Time (sec)

In this image, the input data is represented by the x-axis, while the y-axis shows the total task time in seconds. Blue lines show the total task time taken by Yarn scheduler, whereas red lines show the total task time taken by Fair scheduler. The graph above demonstrates that Fair scheduler outperforms Capacity Yarn scheduler in terms of results for each Hadoop application. The results provided by both the scheduler for a range of Hadoop applications are shown in Figures 4 to 8. Each application receives an effective response from Fair Scheduler. In order to determine the value of the Assessment parameter, all further evaluation will be done on the Fair scheduler.

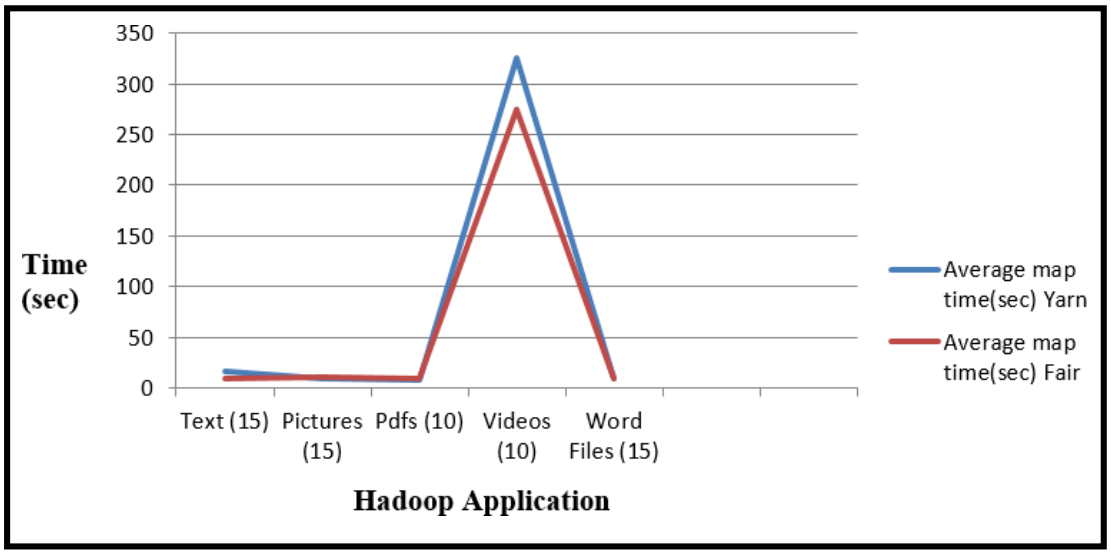

Figure 4: Capacity Analysis YARN vs. Fair Scheduler Average map Time (sec)

In this diagram x-axis represents input data and y- axis represent Map time in seconds and blue line indicates time taken by Yarn scheduler and red line indicates time taken by Fair Scheduler and find that Fair scheduler is more effective as compare to Yarn scheduler in terms of Map time.

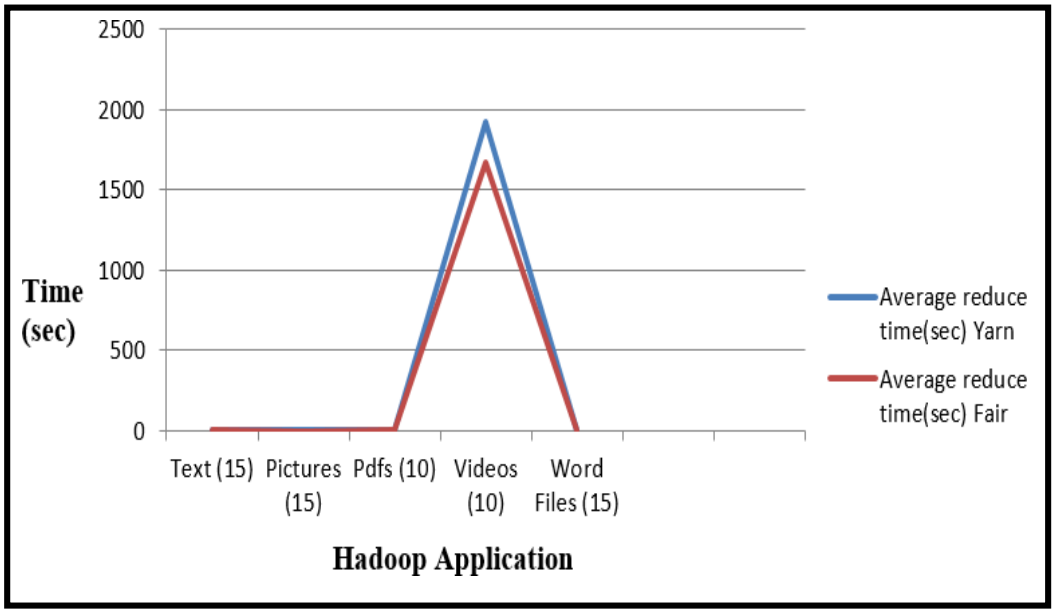

Figure 5: Capacity Analysis YARN vs. Fair Scheduler Average reduce Time

In this diagram x-axis represents input data and y axis represent Reduce time in seconds and blue line indicates time taken by Yarn scheduler and red line indicates Reduce time taken by Fair Scheduler and finds that Fair scheduler is more effective as compare to Yarn scheduler in terms of Reduce time.

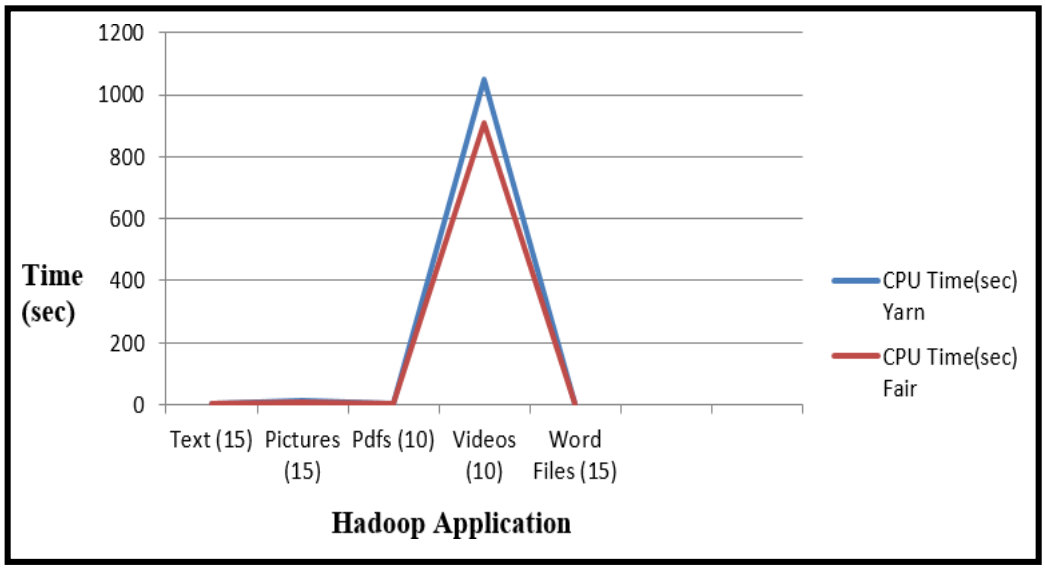

Figure 6: Capacity Analysis YARN vs. Fair Scheduler CPU Time (sec)

In this diagram x-axis represents input data and y axis represent CPU time in seconds and blue line indicates time taken by Yarn scheduler and red line indicates CPU time taken by Fair Scheduler and finds that Fair scheduler is more effective as compare to Yarn scheduler in terms of CPU time also.

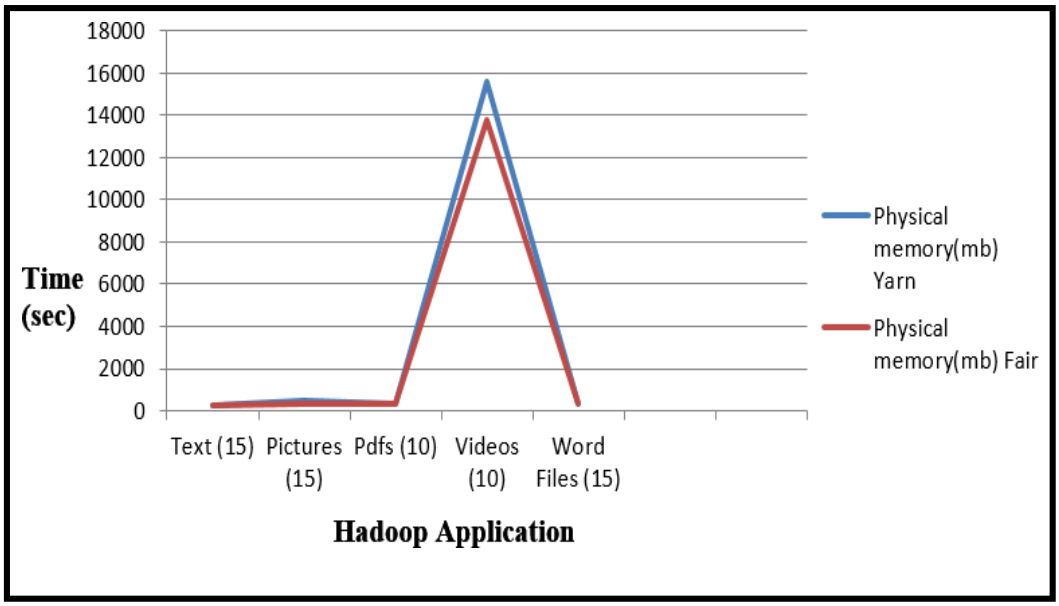

Figure 7: Capacity Analysis YARN vs. Fair Scheduler Physical Memory In MB

In this diagram, the x-axis represents input data, the y-axis, Physical Memory in (MB), and the blue and red lines, respectively, represent Physical Memory taken by Yarn scheduler and Fair scheduler. It is discovered that Fair scheduler is more efficient than Yarn scheduler in terms of Physical Memory.

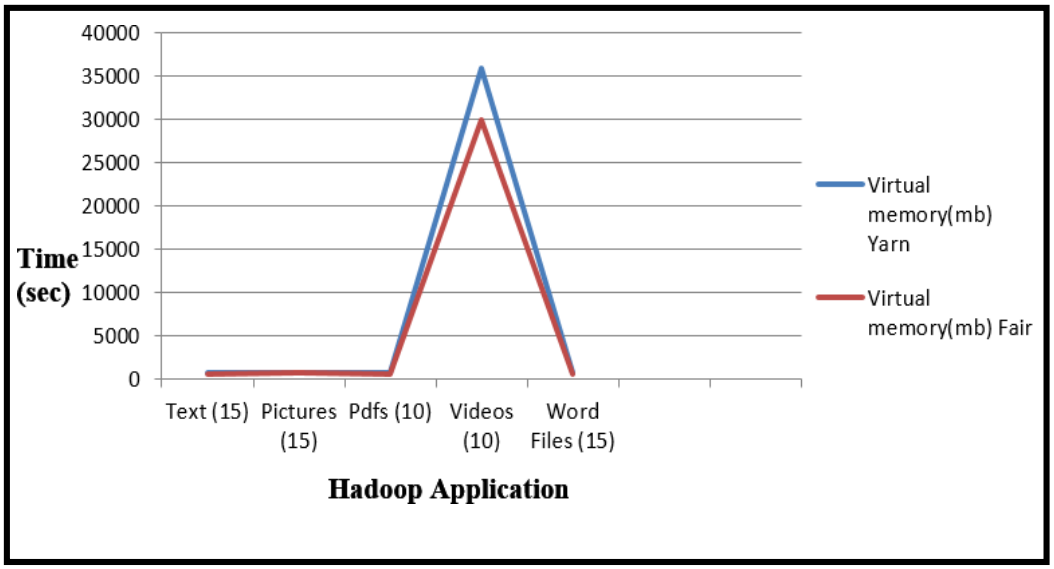

Figure 8: Capacity Analysis YARN vs. Fair Scheduler Virtual Memory In MB

In this picture, the x-axis represents input data, the y-axis represents virtual memory, and the blue and red lines, respectively, show the virtual memory used by Yarn and Fair schedulers. It is discovered that Fair scheduler is more efficient than Yarn scheduler in terms of virtual memory.
Figures 3–8 demonstrate that Fair Scheduler outperforms Capacity Analysis YARN scheduler in terms of performance.

The MEAB and Fair schedules are run and assessed.

Evaluated Parameter:

1) Time Complexity (Encryption Execution Evaluation Time):

To calculate encryption time, five types of files i.e., image file, word processing files, text files, pdfs files and video files are taken and results are presented in following table and graphical form.

Table 6: Encryption Execution Evaluation Time with MEAB

File Type	**Encryption Execution Evaluation Time (Sec)MEAB Algorithm**
Image Files (15)	0.255
Word Processing Files (15)	0.195
Text File (15)	0.152
Pdfs (10)	0.295
Video File (10)	0.415

Table 6 shows evaluation of Encryption Execution Evaluation Time (Sec) of MEAB Algorithm.
Evaluation time is taken in sec and real time input files are taken in MB.

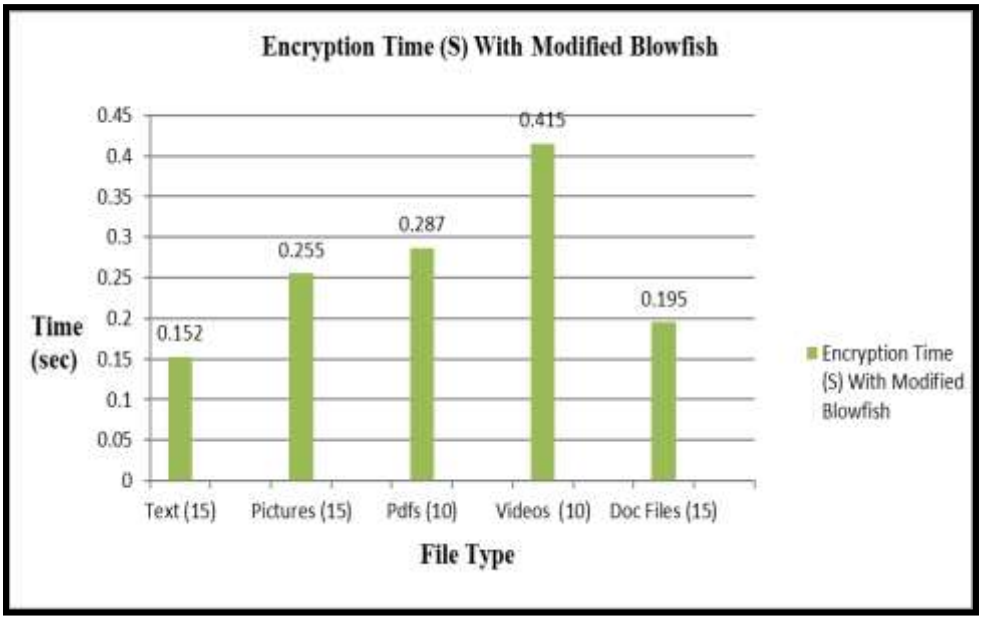

Figure 9: Encryption Execution Evaluation Time with MEAB

Figure 4.7x-axis represents variety of data and y axis represent encryptiontime in seconds using Fair Scheduler.

Variety of data in MB taken to encrypt and decrypt with MEAB. Evaluated values represents in graph and tables

2) Size:

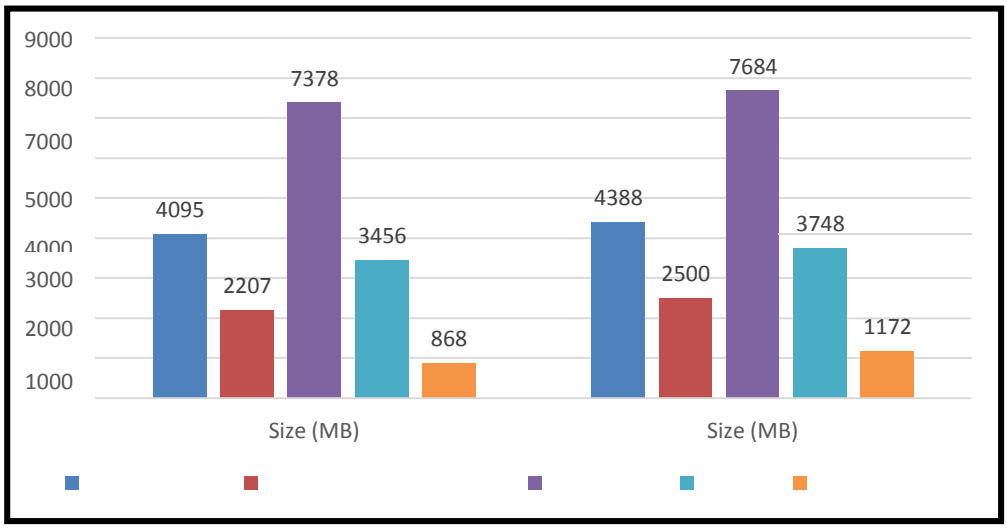

Figure 10: Encrypted Files and original files in MB with MEAB

Figure 4.8 x-axis represents variety of data in MB and y-axis represents Size of the data in MB. Table 7 shows Encryption and decryption of variety of files in MB.

Table 7: Encryption/Decryption of Variety of Files

Files	Size (MB)	EncryptedFiles	Size (MB)	DecryptedFiles	Size (MB)
Text (15)	4095	TestFile-1p	4388	TestFil-1n	4095
Images (15)	2207	TestFile-2p	2500	TestFil-2n	2207
Pdf (10)	7378	TestFile-3p	7684	TestFil-3n	7378
Videos (10)	3456	TestFile-4p	3748	TestFil-4n	3456
Word Files(15)	868	TestFile-5p	1172	TestFil-5n	868

3) Entropy:
To calculate the entropy of files different types files is taken and entropy is calculated in Crypt Tool 2.2. Results presents in table and graph.

Table 7: Entropy of Encrypted Files

Encrypted Files	Size (MB)	Entropy
Image Files (15)	4388	7.86
Word Processing Files (15)	2500	7.77
Text File (15)	7684	7.93
Pdfs (10)	3748	7.86
Video File (10)	1172	7.41

Table 4.6 shows the evaluated Entropy of variety of Encrypted Files and Figure 4.9 shown graphical presentation of entropy of encrypted files with MEAB. MEAB achieves maximum 7.93 entropy value.

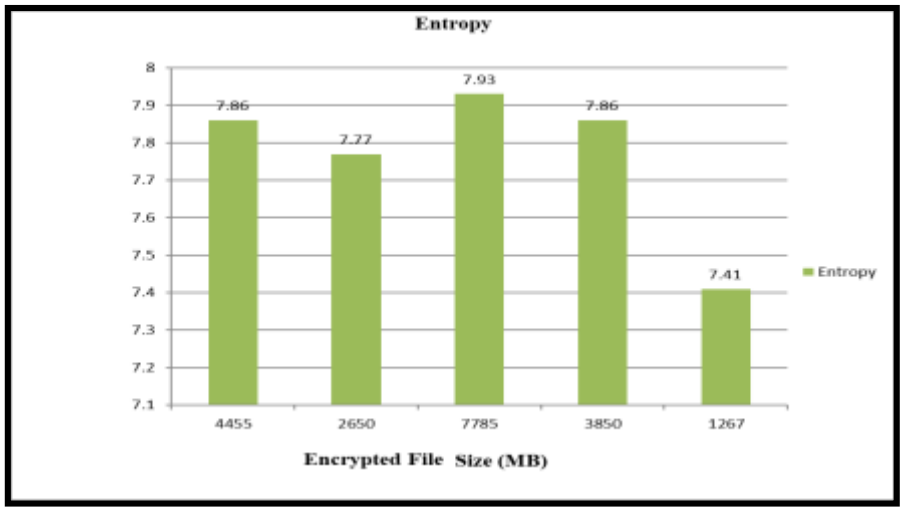

Figure 11: Entropy of Encrypted files

4. CONCLUSION

Electronic assaults, which range from jamming attacks to data breaches brought on by cyber security flaws, pose a threat to businesses, governments, and individuals. Whether carried out by hacktivist groups or state-sponsored cyberwarfare operations, these types of attacks are becoming increasingly frequent. By staying informed on the most recent cyber attacks, businesses and individuals can stay protected. Through The Daily Swig. Businesses were obliged to adapt to new working conditions as a result of the epidemic, which gave hackers new chances. Ransomware is regarded as the most concerning issue in the present climate. The average downtime for organisations that had been attacked in the second quarter of 2021 was 23 days. A Ransomware attack on a corporate network occurred around every eleven seconds in 2021. The proposed enhanced Blowfish algorithm requires less storage than the Blowfish algorithm. User-provided passwords are used to generate keys rather than retaining them. The password entered by the user is converted to UTF-16 format. The password is stored in a key file. This file can be used by modified blowfish Crypt. This tactic's major objective is to render user data incomprehensible to all parties except the receiver. Documentation is required for each map reduce operation that modifies the records. Information about users who are responsible for specific tasks must be documented. Keys stored sideways with encoded data require more capacity, and the key may be easily extracted and the statistics easily decrypted. Passwords are something that most people are familiar with, hence a technique has been developed for producing strong cryptographic keys that only employ passwords that are attainable by humans.
Since protecting information integrity might be difficult, we always work to improve so that we can offer safer encryption procedures. Saving all records from the remote server and guaranteeing integrity is not practical due to the significant quantity of transmission and computation required. As a result, "blockless verification" is created, allowing just metadata—which is built by the CSP and verified at the client end—to be downloaded from the cloud. Data integrity is completed on the client side through a GUI called GtkHash. Before sending the file to the server, a combination of hashing methods including MD5, SHA1, and SHA256 are employed to ensure data integrity. The checksums are updated when a file is downloaded. Integrity is checked using 26474 MB of material, which includes text, images, PDFs, videos, and word files. All files are unlocked using Modified Blowfish. To make sure that all files are secure, GtkHash is employed. Encryption time, key hardness, and HFDS hardening index are used to evaluate the modified blowfish algorithm's efficiency. The proposed Modified has achieved 0.58 % reduction in encryption time compared to AES for text files, 0.19 % reduction in encryption time compared to Blowfish for text files, 3 % reduction in encryption time compared to AES for pictures, 2.29 % reduction in encryption time compared to Blowfish for pictures, 1.9 % reduction in encryption time compared to AES for Pdfs, 1.66 % reduction in encryption time compared to Blowfish for Pdfs, 6.7 % reduction in encryption time compared to AES for videos, 6.10 % reduction in encryption time compared to Blowfish for videos, 3.94 % reduction in encryption time compared to AES for word files, 2.9 % reduction in encryption time compared to Blowfish for word files. The primary techniques that can be applied to further improve results are as follows: Nature Inspired Algorithms , Swarm Intelligence , Support Vector Machines and Metaheuristics.

5. REFERENCES

1. A. Ashabi, S. Bin Sahibuddin, and M. S. Haghighi, "Big Data: Current Challenges and Future Scope," ISCAIE 2020 - IEEE 10th Symp. Comput. Appl. Ind. Electron., pp. 131–134, 2020, doi: 10.1109/ISCAIE47305.2020.9108826.
2. D. Chong and H. Shi, "Big data analytics: a literature review," J. Manag. Anal., vol. 2, no. 3, pp. 175–201, 2015, doi: 10.1080/23270012.2015.1082449.
3. I. Ahmad, H. Bakht, and U. Mohan, "Cloud Computing – A Comprehensive Definition," J. Comput. Manag. Stud., vol. 1, no. 1, pp. 30–2017, 2017.
4. "Knowledge Definition & Meaning - Merriam-Webster," Merriam-Webster. 2022, [Online]. Available: https://www.merriam-webster.com/dictionary/knowledge.
5. N. Kumar, J. Thakur, and A. Kalia, "Performance Analysis of Symmetric Key Cryptography Algorithms: DES, AES and Blowfish," Anu Books, vol. 1, no. 2, pp. 28–37, 2011, [Online]. Available: www.tropsoft.com.
6. S. Ryan, "Wisdom (Stanford Encyclopedia of Philosophy)," Stanford Encyclopedia of Philosophy. pp. 1–8, 2020.
7. N. Chaudhari and S. Srivastava, "Big data security issues and challenges," Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2016, no. May, pp. 60–64, 2017, doi: 10.1109/CCAA.2016.7813690.
8. M. A. U. D. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," Proc. 2014 Zo. 1 Conf. Am. Soc. Eng. Educ. - "Engineering Educ. Ind. Involv. Interdiscip. Trends", ASEE Zo. 1 2014, 2014, doi: 10.1109/ASEEZone1.2014.6820689.
9. G. Priyanka and A. M. Lal, "A hybrid encryption method handling big data vulnerabilities," Int. J. Cloud Comput., vol. 8, no. 3, pp. 207–213, 2019, doi: 10.1504/IJCC.2019.103879.
10. P. Oberoi, "Design a Framework to Detect Malicious Insider Attack in Cloud Based Environment Doctor of Philosophy (Computer Science & Applications)," p. 2019, 2019.
11. J. Thakur and N. Kumar, "DES, AES and Blowfish: Symmetric key cryptography algorithms simulation based performance analysis," Int. J. Emerg. Technol. Adv. Eng., vol. 1, no. 2, pp. 6–12, 2011.
12. A. H. Mahmud, B. W. Angga, Tommy, A. E. Marwan, and R. Siregar, "Performance analysis of AES-Blowfish hybrid algorithm for security of patient medical record data," J. Phys. Conf. Ser., vol. 1007, no. 1, 2018, doi: 10.1088/1742-6596/1007/1/012018.
13. M. Nitesh and M. Rakshak, "Study On Big Data Security Issues for Networking," Int. J. Sci. Res. Comput. Sci. Appl. Manag. Stud. (IJSRCSAMS), ISSN 2319-1953, vol. 7, no. 6, pp. 6–8, 2018.
14. D. K. Chauhan, "A Review of Issues and Challenges with Big Data," no. July, 2017.
15. B. Matt, Introduction to Computer Security, vol. 91. 2017.
16. A. Mathur, "A Research paper : An ASCII value based data encryption algorithm and its comparison with other symmetric data encryption algorithms," Int. J. Comput. Sci. Eng., vol. 4, no. 9, pp. 1650–1657, 2012.
17. P. Oberoi and S. Mittal, Review of CIDS and techniques of detection of malicious insiders in cloud-based environment, vol. 729. 2018.
18. Vormetric Insider Report, "Trends and Future Directions in Data

Security GLOBAL EDITION," 2015. [Online].
19. P. Oberoi, S. Mittal, and R. Kumar, "ARCN: Authenticated Routing on Cloud Network to Mitigate Insider Attacks on IAAS , unpublished."
20. J. Ni, K. Zhang, X. Lin, Q. Xia, and X. S. Shen, "Privacy-preserving mobile crowdsensing for located-based applications," IEEE Int. Conf. Commun., pp. 1–6, 2017, doi: 10.1109/ICC.2017.7997116.
21. B. Padmavathi and S. R. Kumari, "A Survey on Performance Analysis of DES, AES and RSA Algorithm along with LSB Substitution Technique," Int. J. Sci. Res., vol. 2, no. 4, pp. 2319–7064, 2013.
22. G. Kapil, A. Agrawal, and R. A. Khan, "Big Data Security challenges : Hadoop Perspective," Int. J. Pure Appl. Math., vol. 120, no. April, pp. 11767–11784, 2018.
23. A. M. A. Al-Neaimi and R. F. Hassan, "New Approach for Modifying Blowfish Algorithm by Using Multiple Keys," IJCSNS, March, vol. 11, no. 3, pp. 21–26, 2011.
24. G. Manikandan, N. Sairam, and M. Kamarasan, "A new approach for improving data security using iterative blowfish algorithm," Res. J. Appl. Sci. Eng. Technol., vol. 4, no. 6, pp. 603–607, 2012.
25. G. S. Sandhu, V. Verma, and K. Rajesh, "Comparing Popular Symmetric Key Algorithms Using Various Performance Metrics," Int. J. Adv. Res. Comput. Sci. Manag. Stud. ISSN 2321-7782, vol. 1, no. 7, pp. 394–399, 2013.
26. M. Marwaha, A. Singh, and T. Singh, "Comparative analysis of cryptographic algorithms," Int. J. Adv. Eng. Technol. E-issn 0976-3945, vol. 4, no. September, pp. 2013–2016, 2018.
27. J. Juremi, R. Mahmod, S. Sulaiman, and J. Ramli, "Enhancing Advanced Encryption Standard S-Box Generation Based on Round Key," Int. J. Cyber-Security Digit. Forensics, vol. 1, no. 3, pp. 183–188, 2012.
28. N. Kirubanandasarathy, N. Vasudevan, and R. Sarojini, "Vlsi Design and Implementation of pipelined advaced encryption standard architecture," Int. J. Appl. Eng. Res. ISSN 0973-4562, vol. 10, no. 55, pp. 850–853, 2006.
29. A. Patil and R. Goudar, "A Comparative Survey Of Symmetric Encryption Techniques For Wireless Devices," Int. J. Sci. Technol. Res., vol. 2, no. 8, pp. 61–65, 2013.
30. Sumitra, "Comparative Analysis of AES and DES security," Int. J. Sci. Res. Publ., vol. 3, no. 1, pp. 3–7, 2013.
31. L. Ismil, G. Galal-Edeen, S. Khattab, and M. Bahtity, "Performance Examination of AES Encryption," Int. J. Rev. Comput. ISSN 2076-3328, vol. 12, no. December, pp. 18–29, 2012.
32. S. K. Parmar and P. K. C. Dave, "A Review on Various Most Common Symmetric Encryptions Algorithms," vol. 1, no. 4, pp. 2–5, 2013.
33. S. W. Kareem, R. Z. Yousif, and S. M. J. Abdalwahid, "An approach for enhancing data confidentiality in hadoop," Indones. J. Electr. Eng. Comput. Sci., vol. 20, no. 3, pp. 1547–1555, 2020, doi: 10.11591/ijeecs.v20.i3.pp1547-1555.
34. P. Kumar and P. Sharma, "Cryptographic Security technique in Hadoop for Data Security," vol. 6, no. 3, pp. 189–193, 2018.
35. R. B. D. Sharifnawaj Y. Inamdar, Ajit H. Jadhav and A. A. G. Pravin S. Shinde, Indrajeet M. Ghadage, "Data security in Hadoop distributed file system," Int. Res. J. Eng. Technol., vol. 3, no. 4, pp. 939–944, 2016, doi: 10.1109/ICETT.2016.7873697.

36. R. R. Parmar, S. Roy, D. Bhattacharyya, S. K. Bandyopadhyay, and T. H. Kim, "Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions," IEEE Access, vol. 5, no. October 2018, pp. 7156–7163, 2017, doi: 10.1109/ACCESS.2017.2700228.
37. M. Usama and N. Zakaria, "Chaos-based simultaneous compression and encryption for Hadoop," PLoS One, vol. 12, no. 1, 2017, doi: 10.1371/journal.pone.0168207.
38. G. P. Patro, "A Novel Approach for Data Encryption in Hadoop A Novel Approach for Data Encryption in Hadoop."
39. V. N. Patil, "Securing Hadoop using OAuth 2 . 0 and Real Time Encryption Algorithm," no. July, 2015.