A Framework for Secure Data Storage and Retrieval in Cloud Environment

Anita V. Mithapalli, Swati S. Joshi

Abstract: ***Plenty of research work is going on for efficient storage, processing, and analysis of large volume of data generated in real time and having varying nature and quality. The most common open-source framework for efficient computation of such large volume of data is Hadoop which processes big data sets by employing clusters of networked computers. On the other hand, cloud computing refers to storage of data and applications in cloud servers and accessing of the data of applications over the Internet following an on demand scheme. So the organizations who want to reduce costs and complexities associated with big data framework, the most suitable option for them is to take help of cloud infrastructure. But one biggest concern in this regard is the security of data and applications in cloud. Though Hadoop provides in-built encryption scheme and secured HTTP protocol, once data and applications are stored in public cloud, they become vulnerable to various security breaches still remain uncontrolled by the cloud service providers giving rise of a feeling of untrust. In this scenario, encrypting sensitive business data before cloud uploading may help in preventing access of data by evil intruders. In this paper, an extension to Hadoop security with respect to shared cloud has been proposed by designing a software framework where files are encrypted before uploading to cloud. Security performance of this framework for securing data in storage as well as in transit has been implemented such that without using the framework retrieval of data is not at all possible. Extra layer of security aided by symmetric key cryptographic technique has been proposed which will enhance the security of customers' resources along with the present standard security measures of a cloud system. A software system performs symmetric encryption before transmitting a file of any format to cloud. To access this encrypted file, the same software system has to be used to download and decrypt the file. This paper also investigates the performances of most common symmetric key techniques AES, DES and triple DES cryptography with respect to the successful encryption of the customer data. This software framework can be applied to provide an extra security layer at the client's end for users availing service of the cloud platform.***

Keywords: Cloud, Big Data, Hadoop Security, Symmetric key cryptography

I. INTRODUCTION

To run a successful and profitable business one of the important criteria is serving more number of clients. The more clients are served, the more profit can be earned by the business venture. Additionally, in current situation, the business ventures have a huge scope to reach the clients in the whole world through online mode.

Nowadays every existing as well as startup business ventures try to optimize their business goals by the aid of emerging and dominant technologies like machine assisted learning for automatic prediction, big data analytics for analyzing and discovering unknown patterns in the large volume of structured and unstructured data and cloud based computing for the purpose of optimizing their business goals. So any business venture having a dream to spread itself worldwide should manage and handle plenty of customers' transactions arriving in the business premise in various forms, sources, speed of arrival with varying and unstructured data format and above all in large volume. One of the highly popular and efficient technologies that come into handling of such large volume of data in behalf of business ventures is the big data technology. Apache Hadoop is the most common and efficient open-source big data framework which is highly capable of managing such huge amount of data and related processing [1]. There are two main parts of the core Apache Hadoop: Hadoop Distributed File System and a programming facility for data analysis. The files are partitioned into large sized blocks and actually organized over in various cluster nodes. A cluster is a low-end computational node for the purpose of storage and analysis of data with respect to a dispensed computing framework. Data is duplicated across many cluster nodes for ensuring reliability. More clusters can be added if the business data overgrows. Now, if a business venture builds its own private setup, then substantial cost can be incurred to manage and run the private infrastructure. Now, small to medium sized companies trying to step in the domain of big data computing may not arrange for the highly expensive infrastructure. For them, the most suitable option is to rent cloud infrastructures for storage and analysis of large volume of business data.

Cloud computing is one of such Internet-based emerging technologies which plays a great role in supporting specifically small and medium sized business ventures without much financial investment and maintenance cost. There is a list of benefits that can be achieved by the business ventures if they by incorporate cloud computing facility. For instance, the substantial cost associated with building and operating private infrastructure set up can be saved, cloud based application can be accessed in exchange of nominal pay instead of purchasing software and thereby avoiding the purchasing as well as regular updating cost.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

Anita V. Mithapalli, Department Of Computer Science & Engineering, N.B.N. Sinhagad College of Engineering, Solapur, India.
E-mail: g.priyanjali@gmail.com
Prof. Swati S. Joshi, Department Of Computer Science & Engineering, N.B.N. Sinhagad College of Engineering, Solapur, India.
E-mail: ssjoshi.nbnscoe@gmail.com

In other words, the cloud framework provides simplified management of computing resources, accessing of computing resources from a distant location and a cost effective approach. These benefits of cloud computing have attracted nearly 70% of organizations to outsource their businesses to the cloud environment [2].

In spite of so many attracting benefits of cloud computing, every organization must be aware of potential privacy and security threats. After so many organizations are adopting cloud, still it has not been possible to tame and control cloud security threats; it is actually rising continuously with the increasing demands of computing resources of the clients. In a most recent incident, nearly 100 million US customers of Capital One Financial Corp., a credit card issuing consumer banking organization, were exposed by a cloud security breach when their credit card applications containing Social Security Number and bank account number were stolen from an Amazon cloud server [3].

Hence, when such a giant company like Amazon can be vulnerable to cloud security threats, the consumers of cloud facility can never be a position of absolute trust to his cloud provider. As a consequence, the consumers should try to ensure the security and privacy of their business data and application by imposing additional security layer from their side by the aid of some cryptographic techniques such that even though the data is stolen, it can’t be accessed without invading data encryption.

Therefore, this research work focuses on implementing an additional client side security layer for the purpose of protecting client data by the aid of cryptographic algorithms.

This paper is organized into following sections. Section II focuses on previous research works on various threats to cloud security and methods of handling, section III presents a brief overview of cloud technology and most critical security threats, section IV briefly elaborates various cryptographic algorithms applied for data privacy in cloud, section V presents the methodology of implementing client side security, section VI analyses the results and the final section VII concludes the paper with a discussion of pros and cons of implementation and future scope.

II. LITERATURE REVIEW

As the cloud computing services are provisioned over Internet, the traditional security attacks like accessing unauthorized data content by abducting data packets transmitted by others in the network called eavesdropping or the attack where an intruder prevents an authorized user to access the service, known as denial of service are also applicable to cloud network. Here, the focus is on most critical threats to cloud security that can have a great impact on a cloud service consumer.

According to the most recent report published by Cloud Security Alliance (CSA), the eleven most critical threats to cloud computing security are as below [5]. This list has been prepared by considering the opinions of various levels of cloud experts.

- Data Breaches
- Misconfiguration and Inadequate Change Control
- Lack of Cloud Security Architecture and Strategy
- Insufficient Identity, Credential, Access and Key Management
- Account Hijacking
- Insider Threat
- Insecure Interfaces and APIs
- Weak Control Plane
- Metastructure and Applistructure Failures
- Limited Cloud Usage Visibility
- Abuse and Nefarious Use of Cloud Services

The above list has been arranged with respect to the significance of the threats.

According to the above list the most significant concern of the consumers is data breach which refers to viewing, stealing, processing or using confidential and sensitive consumer data by unauthorized intruders [6][7][8].There are various incidents of data breach that occurred in the recent past and still happening make the cloud computing system vulnerable and has become a matter of significant concern. Data breach can result in bad impacts on the goodwill of the provider as the mutual trust between the consumer and the provider can be broken. It has been proved to be the most harmful security breach among others. Unsecured data storage, uncontrolled granting of permissions, deactivating logging or monitoring operations, unconstrained handling of ports and services, lack of efficient change management scheme etc. are some of the factors that cause misconfiguration leading to data breaches [9][10]. Errors in configuring a cloud system ca impact the cloud network with respect to security and performance. The security mechanisms in public cloud should be lined up according to the business goals and appropriate security framework and blueprint must be implemented to prevent cyber-attacks which can adversely impact financial status and reputation of an organization. To maintain the proper standard of cloud services it is very important to maintain proper identification management of the users which is a judgmental affair in cloud service [11]. Access to digital files, computer and other physical resources such as servers must be controlled, scanned and protected. By applying the method of encryption with providing different keys based on authorization limitation can be a suitable way of controlling access to sensitive data [12]. To protect sensitive cloud data against data hacking, the most common way is to enforce encryption and proper controlling of cryptographic keys is a key issue in clod security [13]. Lack of regular updating of security keys, password or certificates, absence of efficient identity management scheme capable of handling identity of large number of users, lack of strong username and password and unconstrained access control can lead to data breaches. If a harmful invader gets access to service accounts by stealing username and passwords or other confidential information by luring a user, data breaches and losses can harm consumer’s cloud subscription leading to disruption of cloud service [14].

One of the most significant security attacks that can’t be tamed by securing cloud infrastructure is actually depends on the human intentions [15]. If a person, insider to the cloud service providing organization, having authorized admittance to computer resources, abuse consumer’s data intentionally or unintentionally, can be the cause of data theft resulting business process disruption. Any existing or previous employee, trusted partner can play the role of an insider.

Collection of Software user interfaces and Application Program Interfaces (APIs) developed by the cloud provider for the purpose of permitting the consumers access and use of the cloud services should be secured following an appropriate security mechanism for preventing them from evil or accidental intrusion. Awfully developed, broken or uncovered APIs can have losing outcomes like data theft or data losses [16].

The problems of conventional networking systems can be mitigated by the use of Software Defined Networking (SDN) which is the new paradigm that is being used frequently in the cloud. SDN actually maintains an architecture with three levels termed as control plane, data plane and application plane. Among these levels, the role of control plane is of utmost important as it acts as the centralized manager of SDN. The open interfacing nature of SDN can put it under various security vulnerabilities and that's why appropriate design of control plane is very much necessary [17]. The person in charge of the control plane should be aware of how consumer data would be transmitted in case of data replication and relocation and which points in the data route are of concern for the purpose of forming a strong control plane leading to efficient protection of business data.

An organization opting for cloud service adoption should efficiently view, monitor and analyze the use of cloud services to identify malevolent activity by any employee by aligning use of cloud services to security policies of the organization. Improper use of cloud service can make existing as well as future incoming data vulnerable to abuse, misuse or corruption.

Evil invaders make use of cloud facility to plant malwares or initiate attacks by hiding behind the authorized domain of the provider. This actually permits the invader to use the cloud service for any illegal and harmful objective.

As with other traditional services, big data services can also be implemented in cloud. Two most important components for managing big volume of data in the cloud environment are MapReduce framework and Hadoop which is actually an open source analytics platform for big data with respect to cloud. But security vulnerability of big data in cloud environment remains, as the possibility of data breaches, theft or corruption are still there [18]. The primary features of big data such as volume, variety and veracity impose challenges on securing big data in cloud environment. The scaling nature of big data and their distributed property may be the hindrance in securing big data with the traditional security measures. As the data types in big data vary in nature like structured, unstructured, linked, probabilistic etc., security measure for one type of data may not be suitable for the other type. The property of veracity confirms that the data being handled are safe from untrusted access of the invaders [19].

Various approaches have been followed by the researchers to secure big data in cloud environment. In a research, the researchers have presented an innovative framework where datacenters are divided in a series of n components and portion of data are stored in some physical device of m cloud storage providers [20]. Instead of encrypting the large sized data, this framework encrypts the path to the data storage. Amazon cloud service has been used for efficient key management and log file processing. Algorithms have been proposed for enhancing cloud security by improving big data privacy [21]. A modified Advanced Encryption Standard (AES) technique has been developed using 256 bits/16 bytes key and provides great level of data security in comparison to standard AES encryption. Researchers also have tried to combine symmetric and asymmetric encryption algorithms for enforcing security of Hadoop based data in cloud environment [22]. Symmetric encryption algorithm has been aided with image as secret key and RSA has been used as asymmetric encryption technique. It has been demonstrated by the authors that the suggested encryption technique enhances data security.

Optimization of data storage and imposing of big sensor data privacy in cloud environment have been implemented in [23]. To avoid insider threats and data vulnerability, a combination of tokenization and mining techniques have been employed. The proposed architecture most suitably performs for structured data and can avoid the danger of compromised token data.

In another research work, a unique security framework has been developed for remotely located cloud data centers by extending MapReduce framework which is termed as G-Hadoop [24]. One cluster Hadoop mechanisms like authentication of user and submission of job have been extended for multiple clustered environments by using Secure Socket Layer (SSL) and encryption mechanisms. Hadoop-based data security over cloud environment has been implemented by the researchers which follows a three level encryption technique [25]. First, Hadoop Distributed File System (HDFS) is secured using Data Encryption Algorithm (DEA), data keys are encrypted using RSA algorithm and private keys of the user are encrypted using International Data Encryption Algorithm (IDEA). Techniques for securing big data over cloud by using classification and encryption methods have been proposed in the literature [26]. For the purpose of protecting data, a modified Naïve Bayes based model of classification has been presented which uses sensitivity level of data to classify. The encrypted sensitive data are placed in separate cloud than that of the cloud to store the non-sensitive data.

A storage framework for big data in cloud computing has been designed for avoiding data damage and loss by following the pattern of master-slave distribution [27]. The burden to the master node is minimized by following the most recent visiting pattern. Balancing the load of data distribution is achieved by storing data into many slave nodes. Big data classification methods also have been developed that classifies big data with respect to whether security is needed or not for that data, that is, sensitive data where security is needed and public data which does not require security [28].

For determining sensitivity of data, the architecture of HDFS MapReduce has been employed. Additionally, an advanced encryption technique has also been followed for securing sensitive files. A cryptographic approach has been followed by researchers which actually try to avoid insider threat by keeping client data out of the reach of the operators working for cloud providers [29]. User files are partitioned and partitioned data are placed in separate servers in the cloud. For protecting Hadoop data clusters in the cloud platform, cryptographic oriented system for controlling access of users to the sensitive data has been experimented by the researchers where the base technique employed is proxy re-encryption [30].

Storing of encrypted data in outside cloud servers makes this system suitable for analyzing big data in cloud. A level of security has been added in the Hadoop framework so that it can continue processing huge amount of data in addition to protecting data also.

All of the above most critical threats listed by CSA can result in theft, loss, corruption or deletion of data and may ruin a business venture as business data about clients, business processes, profit/loss statements, tender details are very much sensitive and confidential resources for running a business. It is also clear that cloud providers have not been able to establish absolute trust of their clients. In addition, securing Hadoop clusters in cloud platform has not been extensively explored by the researchers. Hence, though a client keeps data and application in the cloud, he may enforce additional layer of security specifically on business data such that even data theft occurs, the evil invader may not be able to use the data secured by the client side.

III. UNDERSTANDING OF SOME BASIC TECHNOLOGIES

A. Overview of HDFS

As this paper mainly focus on the Hadoop data security stored in the cloud, a brief architecture of HDFS has been discussed, which is used to store large files.

In HDFS, a large file is partitioned into pre-decided block sizes. Various low-end clusters of one or more than one machine are used to store these blocks. Master/Slave architecture is followed. There is one Master node and the other nodes are Slave nodes.

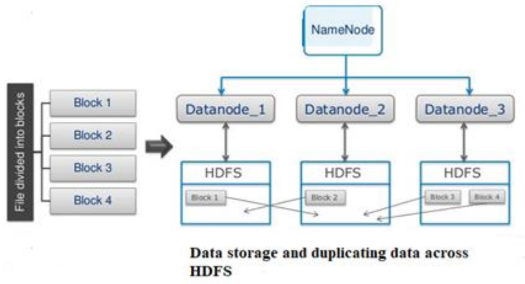


Fig. 1: Overview of HDFS

The NameNode performs access control of files by the clients and responsible for the management of slave nodes as well as data blocks. The detail information about a file like location of the file blocks, size of the files, file permissions, files hierarchy and so on. The DataNodes are basically low-cost and easily available machines. These nodes actually store the data and carry out file read/write requests of the clients. In a regular interval, they convey the detail of the overall state of the HDFS to the NameNode.

B. Overview of Cloud Computing

In simple language, the concept of cloud technology refers to provision of computing services like storage, networking bandwidth, computing servers, computing software, databases, data analytics and machine intelligence by the help of Internet. Cloud computing provides an economical approach of business process outsourcing without any need of specific investment with private infrastructure setup, ubiquitous access of computing resources with flexibility in changing the usage scale. A flexible metered service is provided depending on the customer's usage and a pay-as-you-use model is followed. There are five specific properties of cloud computing that has been defined by The National Institute of Standards and Technology are explained below.

- The customer can himself can allocate computing resources like data storage and network bandwidth according his requirement without any human involvement from the provider's side.
- The customers can access the cloud facility from any type of end systems like a mobile phone, laptop, workstation and tablet phones over the network.
- The computing needs of more than one customer are served at a time by sharing cloud resources of the provider which follows a run time allocation of virtual and physical computing resources.
- With respect to client's point of view, the computing resources can be added or removed according to the client's requirement. In other words, a customer can request and get access to addition computing resources instantly and may release the resources when not required and these events may happen automatically.
- We switch of a light bulb or an electric fan whenever we don't need light or air and bulbs and fans are switched on if we need light or air. The electric current provided to us is metered service. We pay for only that much amount we use. Similarly, cloud providers measure the dynamic resource uses of the clients by a meter like capability for the purpose of efficient control and optimization of the computing resources and the clients pay according to their scaled usage.

A simplified diagram of cloud architecture has been depicted below.

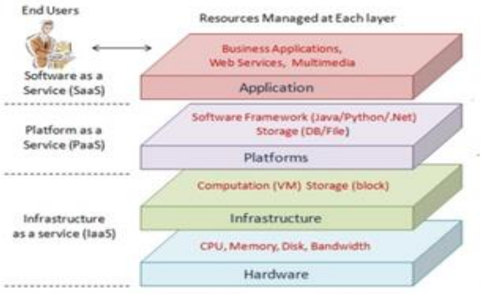

End Users
Resources Managed at Each layer
Business Applications, Web Services, Multimedia
Software as a Service (SaaS)
Application
Software Framework (Java/Python/.Net) Storage (DB/File)
Platform as a Service (PaaS)
Platforms
Computation (VM) Storage (block)
Infrastructure as a service (IaaS)
Infrastructure
CPU, Memory, Disk, Bandwidth
Hardware

Fig. 2: A simplified diagram of cloud environment

According to the above diagram, cloud technology follows service supported architecture. Three most common services are IaaS (Infrastructure as a service), PaaS (Platform as a service) and SaaS (Software as a Service). In IaaS, the computing infrastructure requirement of the client is provided. The infrastructure includes servers, virtual machines, data storage, network bandwidth and operating systems.

The client's requirements for application development, software testing framework, intelligence algorithms, database management systems and more are met by PaaS in addition to servers, data storage and networking bandwidth. Various clod supported applications are provided in SaaS layer such as email and office applications.

C. The Most Suitable Big Data platform: Cloud Computing

The concept of Big Data actually refers to very big volume of real time, structured and unstructured data of varying nature and format which give rise to so much complexity such that this complex data can't be managed by traditional database management system. The primary processing function is to analyze the data and discover unknown information that can be used for optimized decision making. If an organization focuses on working with big data analytics, the first hurdle to cross is to efficiently store this high volume of unstructured data (such as text, images, audio and video) that is being generated in real time manner. The organization can opt for either establishing it's privately owned, on premise infrastructure or go for a cloud storage facility. As the volume of data starts growing exponentially, the organization should also go on scaling it's infrastructure in a progressive manner by more and more investment. Additionally, after managing the complex operations of big data storage, the next operation that are to be performed are processing and analytics of big data which also requires on premise complex software and algorithms to be purchased and maintained. The second approach of moving big data and related applications to cloud is much more feasible as the cloud computing provides scalable storage, efficient processing, cost effective analysis on a secured platform. Various big cloud providers like Google, Microsoft and Amazon offer efficient big data services through their cloud based big data applications. A table below mentions the various cloud based big data applications provided by cloud providers [4] following a pay-as-you-use scheme.

Table 1: Cloud-based big data applications [4]

	Google	Microsoft	Amazon
Big data storage	Google cloud services	Azure	S3
MapReduce	AppEngine	Hadoop on Azure	Elastic MapReduce (Hadoop)
Big data analytics	BigQuery	Hadoop on Azure	Elastic MapReduce (Hadoop)

IV. PROPOSED WORK

Once an organization opts for moving Hadoop clusters to public cloud computing environment, they become prone to various security vulnerabilities as the cloud providers still not fully able to enforce security and privacy of consumer's data. The current Hadoop releases include end-to-end encryption schemes for securing data in storage clusters and when data is travelling over network by the aid of HTTPS protocol. But irrespective of these security measures, there are possibilities of security breach when the clusters are located in cloud premise giving rise to an environment of untrust towards the cloud providers.

Therefore, this research work proposes to extend Hadoop security in the context of shared public cloud. A novel software framework has been designed which encrypts sensitive and confidential data before moving them to cloud infrastructure. A local HDFS instance or a drive has been configured in the shared cloud space and then encrypting data of various formats and varying sizes (large scale encryption) with the same key has been carried out. This sensitive data remain encrypted at cloud environment such that even the URL is shared with others, nobody will be able to access the original content. Without using the software framework, retrieval of this encrypted data from cloud environment is not at all possible. A mechanism is developed where any type of data in various formats (i.e. text, image, audio, video) can be encrypted and put in the directory before uploading to the cloud. The mechanism allows cloud to support the inherent encrypted formats. For example if a jpg file is stored in cloud, as this compression technique changes the bits of original file and may not be supported by cloud. But the proposed mechanism makes it possible.

Decryption of this encrypted data is not carried out in the cloud. Retrieval of the encrypted data comprises of moving this data back to the local host and then performing decryption by the aid of the same software framework used for decryption. Anybody who wants to access the actual data content has to create a local instance of the cloud and provide the same key used for encryption for downloading and accessing the file. As soon as the password is changed at the time of encryption, the data will be no more accessible to the other users.

V. METHODOLOGY

A. Encryption Algorithms Used

Once a consumer outsources his data and application in the cloud, the consumer does not have any access and control over his own data. Although, the consumer ensures his side of security mechanism, once the business data is stored in the cloud environment, the consumer has no permission to the data storage facility. So, the consumer will not be responsible to any data breach at the cloud premise. But, the consumer also must ensure total security of his data as the cloud provider may not be absolutely trusted. The only feasible way to of ensuring data security of the client is to encrypt business data following some encryption algorithms.

There are several ways followed by cloud service consumers for securing data in cloud. Some organizations opt for encrypting their business data before putting in the cloud data storage. In case of any incident of data breach in the provider's end, here, the consumer has at least the satisfaction that he tried to secure his sensitive data. Another way followed by the cloud providers is that whenever data is acquired for storing or transmitting, data encryption is carried out automatically by default. Some organizations fully rely on encrypted connections only without incorporating any specific data encryption algorithm.

In this section, the focus is mainly on securing data by the aid of some encryption algorithm before outsourcing them to cloud storage. Depending on how the key are managed and used, there are two basic approaches followed to encrypt data in cloud: symmetric key cryptography and asymmetric key cryptography. In symmetric key approach one key is used for encryption as well as decryption of user data which is available to both the sender and the receiver. On the other hand, in asymmetric key cryptography, a public, known-to-all key is used for encryption and a private, known-to-only-authorized user key is used for decryption.

Retrieval Number: B3794129219/2019©BEIESP
DOI: 10.35940/ijeat.B3794.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This paper mainly employs three most popular symmetric key algorithms: Triple DES, Advanced Encryption Standard (AES), and RC2.

i) Data Encryption Standard (DES): The DES algorithm was developed by IBM in 1970, which accepts a plain text bit string of predetermined length and following some complex steps converts the plain text bits to an encoded bit string having same length. The fixed length of the input string is 64 bits. An encryption/decryption key is used having 64 bits in size, though only 56 bits are the actual key and the rest of the 8 bits are used for error checking. There are three main stages of the algorithm: Initial permutation (IP), F (Feistel) function and Final permutation (FP). The IP and FP stages are performed once after 64 bit plain text string is input to the algorithm and before the encoded bit string is generated as output. The F stage is iterated for 16 times following same predefined steps. The overall algorithm is stated below.

The 64 bit plain text bit string is taken as input. Then IP is performed on it. The 64 bit permuted string is divided into two halves of 32 bit each. On each of the 32 bit the following operations are carried out. The 32-bit block is extended to 48 bit by employing expansion permutation method. There are eight pieces of 6-bit-each data are generated. From the primary key, sixteen 48 bit subkeys are obtained (16 keys for 16 rounds). The 48-bit expanded block is XORed with a subkey. The outcome from the previous step is separated into eight 6-bit units and forwarded to substitution boxes. Eight 4-bit outputs are obtained using a lookup table. The 32-bit output from the previous step undergoes a fixed permutation and that is the output to the F function. The outcome of the F function from the previous step is fused with the other 32-bit half and a swapping operation is performed on these two halves. Then this 64-bit output of the first round is treated as the input to the next round. Same steps are followed in the next round. After completion of the 16th iteration again a swapping operation is carried out on the 32-bit halves. A final permutation operation is performed on the outcome of the last round and the final result is obtained which is a 64-bit encoded string.

ii) Triple DES: In triple DES a collection of 3 DES keys are employed for encryption/decryption. Three keys, K1, K2 and K3, each having size of 56 bits are used in sequence. The 8-bit error checking is not included in 56 bit size of the keys. First, the 64-bit input plain text bit string is encrypted using single DES key K1. Next, outcome of the previous step is decrypted using DES key K2. Lastly, the result of the previous step is encrypted again by DES key K3. The output is the encoded bit string.

iii) Advanced Encryption Standard (AES): The size of the keys in AES can be either 128 bits for 10 rounds operation or 192 bits for 12 rounds operation or 256 bits for 14 round operations. Whatever the scheme is used, except the last iteration, all other iterations are similar. The steps of the algorithm are explained below keeping 128 bit plaintext input into mind.

• 128 bits plain text is taken as input and divided into sixteen 8-bit units each. These 16 units are arranges in a 4×4 matrix called state, which is actually an array of bytes following column major order. For n rounds in the algorithm, n+1 number of round keys is obtained from the cipher key. The additional key is added with the state matrix initially before entering into the round. A bitwise-OR operation is carried out between a round key block and every 8-bit units of the state matrix. This step is called Add Round Key. In each of the n-1 iterations, four operations are performed, namely, Sub Bytes, Shift Rows, Mix Columns and Add Round Key. Each of these sub steps are explained below.

1. Sub Bytes: Each state matrix byte is substituted by another byte value following a 256 value look up table. The looks up table entries are computed using some mathematical formula enabling in defending mathematical attack.

2. In Shift Rows sub step, the first row of the state matrix is not shifted, but second, third and fourth rows are shifted by one, two and three bytes respectively. The main aim of this step is to create diffusion.

3. Mix Columns: By performing the operation of matrix multiplication, the columns of the state matrix are processed and a new set of columns are generated which substitutes the old ones.

4. Add Round Key is same as previously explained where the bytes of the state matrix are XORed with the appropriate round key and the elements of the state matrix is substituted by the result.

The above sub steps are identical for all the n-1 rounds. In the nth final round, the sub steps that are carried out are Sub Bytes, Shift Rows and Add Round Key. The encoded string is the resultant state matrix.

iv) RC2: RC2 is a symmetric key algorithm developed by Ron Rivest. The algorithm consists of three steps: key expansion, encryption and decryption. The user supplied initial key is and an extended key 64 words is generated where each word is of 16 bits. The 64 bit plain text is placed in 16-bit words R[0], R[1], R[2] and R[3]. Encryption converts the plain text to cipher text at the same place. Decryption follows the opposite process of encryption.

<u>Key expansion</u>

Assuming a byte operation, a key buffer is denoted as L[0], L[1] ,….., L[127]; each L[i] is of 8 bits.

- User inputs T byte key. Where 1<=T<=128. T1 denotes the maximum effective length of the key in n bits
- User supplied key is stored in the key buffer L[i], where i=0, …, , T-1.
- Effective length of the key in bytes denoted by T8 and a mask TM depending on the number of bits in T1 following the formula below.

  T8 = (T1+7)/8;

  TM = 255 MOD 2^(8 + T1 - 8*T8)
- An additional array P[j], where j=0, …, 255, is used for key expansion which contains values from 0 to 255 generated using random permutation.
- Computation for key expansion is as below

```
for i = T to 127 do
        L[i] = P[L[i-1] + L[i-T]];
        L[128-T8] = P[L[128-T8] & TM];
for i = 127-T8 down to 0 do
            L[i] = P[L[i+1] XOR L[i+T8]];
```

The values from L[128-T8] to L[127] is the resultant expanded key.

<u>Encryption</u>

- Encryption consists of operations: “Mix” and “Mash”.

Here, s[0]=1, s[1]=2, d[2]=3, s[3]=4. In “Mix” operation, indices in R[i] are always treated as modulo 4. For example, if i=0, then R[i-1] =R[3]. The operation Mix R[i] is defined as below.

R[i] = R[i] + K[j] + (R[i-1] & R[i-2]) + ((~R[i-1]) & R[i-3]);
j = j + 1;
R[i] = R[i] rol s[i]; where x roll k denotes x undergoes left rotation by k bits.

- Then the operations Mix R[0], Mix R[1], Mix R[2], Mix R[3] form mixing round operation.
- The Mash R[i] operation is carried out as below. R[i] = R[i] + K[R[i-1] & 63];
- Similarly, the operations Mash R[0], Mash R[1], Mash R[2], Mash R[3] form Mash round.

The steps of overall encryption algorithm are mentioned below.

1. Input plain text block of 64 bit is contained in R[0], ..., R[3].
2. Input key is expanded.
3. j=0.
4. Five MIXING rounds are carried out
5. One MASHING round is carried out.
6. Six MIXING rounds are carried out.
7. One MASHING round is carried out.
8. Five MIXING rounds are carried out.
9. Now, R[0], ..., R[3] contain the cipher text.

B. Overall methodology

The steps of overall features have been explained below. Any type of file (text, image, audio and video) and having any size can be encrypted (large scale encryption) before moving them to cloud environment. A software framework resides between the data and the cloud. Any data file to be uploaded to the cloud will go through the software framework which encrypts the files using same key for all files. Decryption process follows the same steps.

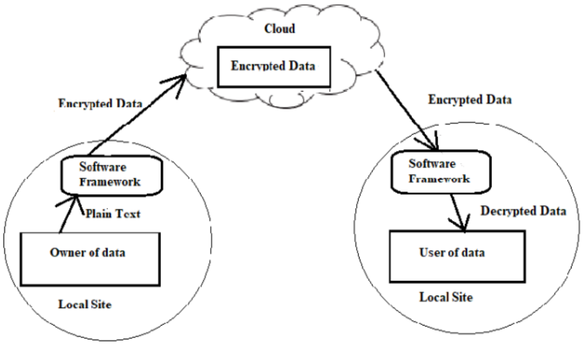


Fig. 3: Encryption and decryption for the proposed work

Microsoft OneDrive has been taken as the cloud computing environment where the encrypted data files are being sent. There is only one user involved who is actually encrypting/uploading and decrypting/downloading data to/from the cloud environment. So there is no question of key sharing over unsecured network.

Step 1: Encryption process

Files to be encrypted and uploaded to the cloud environment are kept in a directory. The developed software framework is executed. The CloudSec window is opened. The Directory menu item is clicked and the directory of files to be encrypted and uploaded is selected. The list of files in the corresponding directory is displayed in a table with the fields like file name, size, last modified time, last back up time etc. Next, the encryption algorithm to be used is chosen. The user clicks on SkyDrive menu item and chooses log in to SkyDrive. OneDrive API Browser opens. In OneDrive API Browser window, the sign in option from file menu is selected. User sign in with Microsoft account. File Service menu item in CloudSec window is selected. A password entering dialog box appears, where user sets his password. The length of the password should be of 8 characters. As the files are being uploaded, their SkyDrive file ids appear in a window and the graphs corresponding to upload time in seconds, encryption time in ms, upload bandwidth in Kbps for the selected encryption algorithm start filling up. The message “uploading files completed” appears after the files have been uploaded to OneDrive. OneDrive API Browser window displays the files uploaded. In this situation, if anybody else has access to the username and password of OneDrive account of the user, he/she can enter into the user’s OneDrive account by opening OneDrive in a web browser but still will not be able to see the actual content of the files uploaded. So, the data files are still encrypted in parking.

Another important thing to be noticed that the files stored in the local drive directory are also encrypted. The user himself can’t access the original contents of the file without using the software framework. This proves that the files were encrypted while in transit.

Step 2: Decryption process

To decrypt and access the original contents of the files, the Download option from File Service menu item in CloudSec window is selected. A password entering dialog box appears. The same password used for encryption and the same encryption algorithm is selected.

In the destined folder, the files are downloaded one after another. As the files are being downloaded from cloud, file ids are displayed in a window. In addition, the graphs corresponding to measure download time in seconds, decryption time in milliseconds and download bandwidth in Kbps are filled up. After completion of the download, “Files downloaded from SkyDrive” appears. Now in the destined folder all decrypted files are residing. As mentioned earlier that local copies of these files are also encrypted. Hence to access the original files, the local encrypted files should be replaced by decrypted and downloaded files in the destined directory. For an uploading/downloading session, the upload/download directory will remain fixed.

Step 3: Performance comparison of encryption algorithms

To compare the performance of symmetric cryptography algorithms, a collection of graphs are displayed in the right portion of the Data Security window. With respect to each algorithm, the following factors have been measured at the time of encryption/uploading and decryption/downloading.

a) Upload time in seconds b) Encryption time in seconds
c) Upload bandwidth in Kbps d) Download time in seconds
e) Decryption time in seconds f) Download bandwidth in Kbps

VI. RESULTS AND DISCUSSION

A software framework has been developed which is used to upload and download directory files from the local client cloud node by a single user using three symmetric encryption methods AES, Triple DES and RC2. In one session for uploading/downloading and encryption/decryption only one local directory has to be considered. The primary purpose is the protect client data in transit as well as in storage when the client is accessing third party cloud service.

For experimental purpose various types of file like text, image, audio, video etc. are taken from the local system. Any of the encryption techniques is selected among AES, Triple DES or RC2. The files are uploaded using the selected encryption technique.

A. Experiment 1: To Prove Safety of the Files in Transit As Well As In Parking

To prove the above, the files in the OneDrive cloud are tried to be accessed. But it is found that unauthorized access of the files in the cloud storage is not allowed. On opening the files in the OneDrive cloud only binary content is shown.

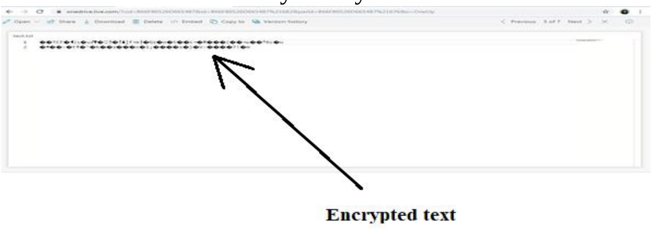


Fig. 4: Encrypted file content in OneDrive displayed by unauthorized user access

This establishes the fact that the files in the third party cloud storage is secured.

Now, the files in the local storage are accessed. It is observed that the files in the local storage are also encrypted and secured and can't be accessed by any unauthorized users. It also proves that the files when transmitted to the cloud were encrypted and are protected against attacks on transit and data integrity is maintained. After downloading the encrypted files from the cloud storage, they are automatically decrypted by the software framework. Now, the last important thing is to replace the locally stored encrypted files with the downloaded and decrypted files.

B. Experiment 2: Performance Analysis of Each Encrypted Methods

Each of AES, Triple DES and RC2 encryption techniques are analyzed with respect to the metrics of upload/download time in seconds, encryption/decryption time in seconds and upload/download bandwidth in Kbps. The encryption techniques are also analyzed with respect to encryption difference. The corresponding graphs have been displayed in Figure 5. For each graph at the right pane, the X-axis denotes number of files that are being uploaded and the corresponding Y-axes are upload time in seconds, encryption time in ms and upload bandwidth in Kbps for both uploading/downloading graphs. For the encryption difference graph X-axis is the number of files and Y-axis is byte differences between encrypted and the original files. If the byte difference is more, the corresponding encryption algorithm is better.

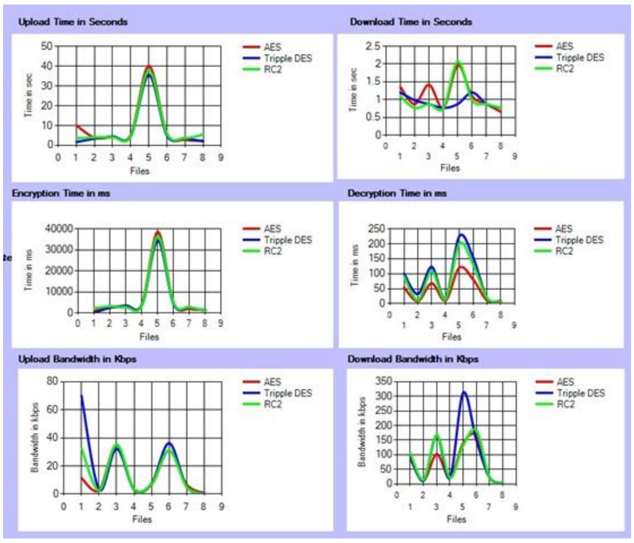


Fig. 5: Performance comparison graphs of the three encryption techniques

Let us discuss the encryption algorithms with respect to the metrics considered.

a) Upload time: Encryption changes the file size. But as the files are being transmitted over network, file sizes must be kept as minimum as possible for faster uploading. According to the graph generated, Triple DES and RC2 perform better than AES.

b) Encryption time: The less is the encryption time, the more efficiently the encryption technique performs. In this case also Triple DES and RC2 performance is better.

c) Upload bandwidth: For better performance, the bandwidth used should be minimum. Here, the AES and RC2 function more efficiently.

d) Download time: For efficient functioning, the encryption algorithm should take less time to download files. In this regard, Triple DES perform well as compare to AES and RC2

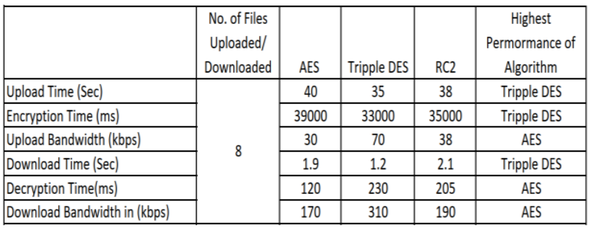
e) Decryption time: The less the time taken to decrypt files, the more efficient the corresponding encryption algorithm is. AES has better performance than Triple DES and RC2 in this case.

f) Download bandwidth: Download bandwidth should be less for optimized performance. AES and RC2 perform more efficiently.

By aggregating the individual performances of each encryption techniques, the overall performance can be concluded as below.

The overall performance of AES and Triple DES is better than RC2. The corresponding performance table of the three encryption algorithms has been given in Table 2 below.

Table 2: Encryption algorithm performance comparison



	No. of Files Uploaded/ Downloaded	AES	Tripple DES	RC2	Highest Permormance of Algorithm
Upload Time (Sec)	8	40	35	38	Tripple DES
Encryption Time (ms)		39000	33000	35000	Tripple DES
Upload Bandwidth (kbps)		30	70	38	AES
Download Time (Sec)		1.9	1.2	2.1	Tripple DES
Decryption Time(ms)		120	230	205	AES
Download Bandwidth in (kbps)		170	310	190	AES

Retrieval Number: B3794129219/2019©BEIESP
DOI: 10.35940/ijeat.B3794.129219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

VII. CONCLUSION

When a client takes the service of a third party cloud provider, there is possibility of security breach of client data. That's why there is a need for enforcing some additional security measure in the client side. The security of big volume of data in cloud environment is a very significant issue. In this paper an additional layer of security has been integrated for Hadoop clusters in cloud environment. A single user at the client side uploads/downloads as well as encrypts/decrypts files to the cloud by the help of a software framework which takes the full responsibility to keep the files secured in the cloud as well as in the transit.

Three symmetric encryption techniques namely AES, Triple DES and RC2 have been used for the purpose of encryption and decryption. They have been compared with respect to different metrics like upload/download time, encryption/decryption time, upload/download bandwidth and encryption difference. It has been observed that among the three encryption techniques AES and Triple DES perform better than RC2 with respect to optimized encryption/decryption as well as encryption difference.

Each of cryptographic algorithms has weakness points and strength points. We select the cryptographic algorithm based on the demands of the application that will be used. From the experiment results and the comparison, the AES algorithm is the perfect choice in case of encryption time and downloads bandwidth. If confidentiality and integrity are major factors, AES algorithm can be selected. And also if the demand of the application is the network bandwidth, the AES is the best option.

REFERENCES

1. https://en.wikipedia.org/wiki/Apache_Hadoop
2. Louis Columbus, "Analytics, Data Storage Will Lead Cloud Adoption In 2017", Forbes, Nov 20, 2016
   Web source: https://www.forbes.com/sites/louiscolumbus/2016/11/20/analytics-data-storage-will-lead-cloud-adoption-in-2017/#74e63c357e7a
3. Robert McMillan, "Capital One Breach Casts Shadow Over Cloud Security", The Wall Street Journal, July 30, 2019
   Web source: https://www.wsj.com/articles/capital-one-breach-casts-shadow-over-cloud-security-11564516541
4. Ibrahim Abaker Targio Hashem et al, "The Rise of Big Data on Cloud Computing: Review and Open Research Issues", Information Systems, Vol. 47, pp. 98-115, 2015
5. " Top Threats to Cloud Computing the Egregious 11", CSA Report, 2019
6. Govind Rao Mettu1 and Dr. Anitha Patil, "Data Breaches as Top Security Concern in Cloud Computing", International Journal of Pure and Applied Mathematics, Vol. 119, No. 14, pp. 19-28, 2018
7. Deba Prasead Mozumder, Md. Julkar Nayeen Mahi, Md Whaiduzzaman, "Cloud Computing Security Breaches and Threats Analysis", International Journal of Scientific & Engineering Research, Vol. 8, Issue 1, pp. 1287-1297, 2017
8. Yogachandran Rahulamathavn, "Assessing Data Breach Risk in Cloud", International Conference on Cloud Computing Technology and Science (CloudCom), 2015, IEEE
9. K. Wood E. Pereira, "International Journal Multimedia and Image Processing (IJMIP)", Vol. 1, Issues 1/2, pp. 17-25, 2011
10. Pericherla Satya Suryateja, "Threats and Vulnerabilities in Cloud Computing", International Journal of Computer Sciences and Engineering, Vol. 6, Issue 3, pp. 297-302, 2018
11. Umme Habiba, Rahat Masood1, Muhammad Awais Shibli and Muaz A Niazi, "Cloud identity management security issues & solutions: a taxonomy", Complex Adaptive Systems Modeling, Vol. 2, Article number 5, pp. 1-37, 2014
12. Enrico Bacis et al, "Access Control Management for Secure Cloud Storage", International Conference on Security and Privacy in Communication Systems SecureComm, pp 353-372, 2016
13. Ramaswamy Chandramouli, Michaela Iorga, Santosh Chokhani, "Cryptographic Key Management Issues & Challenges in Cloud Services", NIST Interagency or Internal Report, 2013
14. Sreenivas Sremath Tirumala, Hira Sathu, Vijay Naidu, "Analysis and Prevention of Account Hijacking based INCIDENTS in Cloud Environment", International Conference on Information Technology (ICIT), 2015, IEEE
15. Atulay Mahajan, Sangeeta Sharma, "The Malicious Insiders Threat in the Cloud", International Journal of Engineering Research and General Science Volume 3, Issue 2, Part 2, pp. 245-256, 2015
16. Betrand Ugorji, Nasser Abouzakhar and John Sapsford, "Cloud Security: A Review of Recent Threats and Solution Models", ICCSM2013-Proceedings of the International Conference on Cloud Security, 2013
17. Deepak Singh Rana, Shiv Ashish Dhondiyal, Sushil Kumar Chamoli, "Software Defined Networking (SDN) Challenges, issues and Solution", International Journal of Computer Sciences and Engineering", Vol. 7, Issue 1, pp. 884-889, 2019
18. Sandeep Kumar Mohapatra, "Cloud Computing and Hadoop Security Analysis", Global Journal of Engineering Science and Researches, pp. 1-5, 2017
19. Yuri Demchenko, Canh Ngo, Cees de Laat, Peter Membrey, Daniil Gordijenko, "Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure", Workshop on Secure Data Management SDM 2013, pp. 76-94, 2013
20. Gunasekaran Manogarana, Chandu Thotab, M. Vijay Kumar, "MetaCloudDataStorage Architecture for Big Data Security in Cloud Computing", International Conference on Recent Trends in Computer Science & Engineering, Procedia Computer Science, Vol. 87, pp. 128 – 133, 2016
21. Christos Stergiou & Kostas E. Psannis, "Efficient and secure BIG data delivery in Cloud Computing", Multimedia Tools and Applications, Vol. 76, Issue 21, pp. 22803–22822, 2017
22. Danish Shehzad, Zakir Khan, Hasan Dağ, Zeki Bozkuş, "A Novel Hybrid Encryption Scheme to Ensure Hadoop Based Cloud Data Security", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 4, 2016
23. Shanto Roy, Ahmedur Rahman Shovon and Md. Whaiduzzaman, "Combined approach of Tokenization and Mining to secure and optimize Big Data in Cloud Storage", IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017
24. Jiaqi Zhao et al, "A security framework in G-Hadoop for big data computing across distributed Cloud data centres", Journal of Computer and System Sciences, Vol. 80, pp. 994–1007, 2014
25. Chao YANG, Weiwei LIN, Mingqi LIU, "A Novel Triple Encryption Scheme for Hadoop-based Cloud Data Security", Fourth International Conference on Emerging Intelligent Data and Web Technologies, 2013
26. Gitanjali, Dr. Kamlesh, "Securing Big Data over Cloud Using Classification and Encryption Techniques", IJRECE Vol. 6, Issue 2, pp. 678-682, 2018
27. Xuebin Chen, Shi Wang, Yanyan Dong, and Xu Wang, "Big Data Storage Architecture Design in Cloud Computing", National Conference on Big Data Technology and Applications, pp. 7-14, 2015
28. Ismail Hababeh, Ammar Gharaibeh, Samer Nofal, and Issa Khalil, "An Integrated Methodology for Big Data Classification and Security for Improving Cloud Systems Data Mobility", IEEE Access, Vol. 7, pp. 9153 – 9163, 2018
29. Yibin Li et al, "Intelligent cryptography approach for secure distributed big data storage in cloud computing", Information Sciences, Vol. 387, pp. 103-115, 2017
30. David Nunez, Isaac Agudo, Javier Lopez, "Delegated Access for Hadoop Clusters in the Cloud", IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2014), pp. 374-379, 2014.

AUTHORS FROFILED

Anita V. Mithapalli, completed her AMIE (Associative Member of Institution of Engineers) examination of Computer Science and Engineering in 2012 from Institution of Engineers India (Kolkata). She is currently doing Master of Engineering in Computer Science and Engineering from N. B. Navale Sinhgad College of Engineering, Kegaon, Solapur affiliated to Punyashlok Ahilyadevi Holkar Solapur University, Solapur. She is currently working as a Technical support staff in Department of Information Technology at Government Polytechnic, Solapur, governed by Directorate of Technical Education, Mumbai since 2013. Her research interest is in security issues in cloud and big data environment, Cloud computing, Hadoop.

Swati S Joshi, has completed her Bachelor of Engineering in 1993 from Walchand Institute of Technolgy affiliated to Shivaji University Kolhapau. And Master of Engineering in Computer Science & Engineering in 2008 from Government College Engineering, Aurangabad. Currently she is working as Assistant Professor in Department of Computer Science and Engineering, N B Navale Sinhgad College of Engineering, Kegaon Solapur affiliated to Punyashlok Ahilyadevi Holkar Solapur University, Solapur. She has 20 years of teaching experience. Her research interests include Artificial Intelligence, Networking, Security, and IoT.