Mehak Int. Journal of Engineering Research and Applications www.ijera.com
ISSN : 2248-9622, Vol. 4, Issue 9(Version 3), September 2014, pp.133-138

RESEARCH ARTICLE OPEN ACCESS

Improving Data Storage Security in Cloud using Hadoop

Mehak *, Gagandeep **
* (Department of Computer Science, Punjabi University, Patiala, India)
**(Department of Computer Science, Punjabi University, Patiala, India)

ABSTRACT
The rising abuse of information stored on large data centres in cloud emphasizes the need to safe guard the data. Despite adopting strict authentication policies for cloud users data while transferred over to secure channel when reaches data centres is vulnerable to numerous attacks .The most widely adoptable methodology is safeguarding the cloud data is through encryption algorithm. Encryption of large data deployed in cloud is actually a time consuming process. For the secure transmission of information AES encryption has been used which provides most secure way to transfer the sensitive information from sender to the intended receiver. The main purpose of using this technique is to make sensitive information unreadable to all other except the receiver. The data thus compressed enables utilization of storage space in cloud environment. It has been augmented with Hadoop's map-reduce paradigm which works in a parallel mode. The experimental results clearly reflect the effectiveness of the methodology to improve the security of data in cloud environment.
Keywords-Cloud, AES, Hadoop, Mapreduce, Data Compression

I. INTRODUCTION

Cloud Computing is a natural step in the evolution of on-demand information technology services and products. To a large extend Cloud Computing is based on virtualized resources. It became popular in October 2007 when IBM and Google announced collaboration in this domain. Cloud Computing is based on Service-Oriented Architecture where the end users request an IT service at a desired functional, quality and capacity level, and receive it either at the time requested or at a specified later time. Virtualization is a pillar for Cloud Computing or it allows abstraction and isolation of lower-level functionalities and underlying hardware and enables portability of higher-level functions and sharing and/or aggregation of physical resources. Cloud Computing security (sometimes referred to simply as "Cloud security") is an evolving sub-domain of computer security, network security, and, more broadly, information security. It refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of Cloud Computing. It refers to the set of procedures, processes and standards designed to provide information security assurance in a Cloud Computing environment. Encryption is the process of encoding messages or information in such a way that only authorized parties can read it. In an encryption scheme, the message or information, referred to as plaintext, is encrypted using an encryption algorithm, generating cipher text that can only be read if decrypted. Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers and for developing distributed applications that can process very large amounts of data. Hadoop parallelizes data processing across many *nodes* (computers) in a *compute cluster*, speeding up large computations and hiding I/O latency through increased concurrency. Hadoop is especially well-suited to large data processing tasks (like searching and indexing) because it can leverage its distributed file system to cheaply and reliably replicate chunks of data to nodes in the cluster, making data available locally on the machine that is processing it. It is a platform that provides both distributed storage and computational capabilities. The distributed storage has the distributed file system called HDFS that provides storage and the computational tier uses a framework called Mapreduce.

The advent of Web 2.0 has made organisations move toward cloud with computing models using Information Technology as a service. The emerging field in the information technology focuses on applications running over the internet such as Saas (Software as a service) or hardware systems available in Data Centres. In both the cases management of data and services rendered by data centres are not fully trustworthy. In the recent years, data storage has enjoyed a period of unprecedented growth. According to hype cycle of big data, the surveys showed the major challenges identified in the recent years. The previous work was there was comparison

between the parallel method and the sequential method. Several security issues have been identified in the current Cloud Era. It is necessary to secure data at rest. Hence, storage encryption is proposed to invoke confidently of large data encryption. Being a time consuming process, this is controlled by an efficient application of the process in parallel mode. In this regard Hadoop's MapReduce and source implementation seems to be an attractive cost efficient solution for processing large scale data. After encryption so as to save the space the compression is done.

II. LITERATURE SURVEY

During the period of time several research papers have been studied which are summarized below:-

A solution for secure storage of the records in a data base by using AES is provided in [1] which is prone to less security attacks and should take optimal time for storage. The key will be rotation mode based on simple logic being implemented in the algorithm thus making it quite complex for attacks.

The proxy servers can convert encrypted files for the owner to the encrypted file for the receiver without the necessarily of knowing the content of original file [2]. These schemes are secure against different types of attacks and all the access permissions made by owner himself. The opportunities for cryptography to address the security challenges in the area of Cloud storage, communication and virtualization is given in [3]. The parallel computing is a promising technique in [5] is analyzed to improve the performance of cryptography and mainly divide and conquer strategy is used in parallel computation and parallel computation gives much better results than sequential computation.

The prevented data access from unauthorized access is discussed in [6] and proposed a scheme that perfectly stores the data and identifies the any temper at the cloud server. It also performs some of the tasks like data updating, deleting, appending.

The data storage correctness issue in reference of cloud computing is analyzed in [7]. They provided an algorithm for trusted and secure data storage model with new encryption scheme and integrity verification. The features of algorithm are useful to reduce computational cost and give almost all the solutions for trusted and secure data storage on cloud.

A scheme 'Two Layer Encryption' means double Encryption for securely outsourcing the data in the cloud was proposed in[8]. The outer layer will be decrypted by the cloud and the inner layer will be decrypted by the user only, by this manner data/information will be highly secured.

A method to build a trusted computing environment for Cloud Computing system by providing secure cross platform into cloud computing system is proposed in [9]. It also provides some important computing security services including authentication, encryption, decryption and compression.

The storage techniques in cloud computing and several storage techniques that provide security to data in cloud with the advantages and drawbacks of these techniques are studied in [10].

NubiSave, a user friendly storage controller implementation with adaptable overheads which runs on and integrates into typical consumer environments as a central part of an overall storage system was introduced in [11]. It also presented a systematic approach for achieving optimality in cloud storage services along the provider's and consumer's iteration of the service life cycle.

Accountable Mapreduce as an additional component for the current Mapreduce model. It is proposed in [12] Accountable Mapreduce employs an auditor group to conduct an A-test on every worker in the system. If malicious behavior occurs the AG is able to detect it and provide verifiable evidence. To improve the evidence they introduce P-Accountability in a-Test to trade the degree of accountability with efficiency.

III. ENCRYPTION TECHNIQUE

The Advanced Encryption Standard (AES) is formal encryption method adopted by the National Institute of Standards and Technology of the US Government, and is accepted worldwide. The AES encryption algorithm is a block cipher that uses an encryption key and a several rounds of encryption. A block cipher is an encryption algorithm that works on a single block of data at a time. In the case of standard AES encryption the clock is 128 bits, 16 bytes in length. The term "rounds" refers to the way in which the encryption algorithm mixes the data re-encrypting it to ten to fourteen times depending on the length of the key. The AES algorithm itself is not a computer program or computer source code. It is a mathematical description of a process of obscuring data. A number of people have created source code implementations of AES encryption, including the original authors.

3.1 Encryption Keys

AES encryption uses a single key as a part of the encryption process. The key can be 128bits (16 bytes), 192 bits (24bytes), 256 bits (32 bytes) in length. The term 128-bits encryption refers to the use of a 128-bit encryption key. With AES both encryption and decryption are performed using the same key. This is called a symmetric encryption algorithm. Encryption algorithm that uses two different keys, a public and a private key, are called asymmetric encryption algorithms.

An encryption key is simply a binary string of data used in an encryption process. The same encryption key is used for encryption and decryption

of information. It is important to keep encryption key a secret and to use the keys that are hard to guess. Some keys are generated by software used for this specific task. Another method is to derive a key from pass phase. Good encryption systems never use a pass phase alone as an encryption key.

3.2 Modes of Operation

There are different methods of using keys with the AES encryption method. These different methods are called "modes of operation". The NIST defines six modes of operation that can be used with AES encryption:

- Electric code book(ECB)
- Cipher block chaining(CBC)
- Counter(CTR)
- Cipher feedback(CFB)
- Output feedback(OFB)
- Galois Counter Mode(GCM)

Each mode uses AES in a different way. For example, ECB encrypts each block of data independently. CTR mode encrypts a 128- bit counter and then adds that value to the data to encrypt it. CBC mode uses an initialization vector and adds encrypted value of each block to the data in the next block before encrypting it. Some modes require you to only encrypt data that is a multiple of 16 byte block size; others allow you to truncate unused data

IV.HADOOP'S MAPREDUCE PARADIGM

Google proposed Mapreduce to simplify data processing on large clusters. Hadoop is the most popular open source implementation of MapReduce framework developed by yahoo and apache software foundation. Mapreduced is deployed as a powerful data processing service over open source systems and has become increasingly popular for its parallel programming framework. Hadoop's MapReduce seems to be an attractive cost effective solution for large scale data processing services like securing the data in cloud through block encryption. Mapreduce can fit in any kind of environment like closed or open systems. The frame work is designed for writing applications that rapidly process vast data during runtime on compute clusters. The code automatically partitions input data performs scheduling, monitoring and performs fault tolerance mechanism through which it re-executes the failed tasks in the commodity of large cluster machines.

Hadoop is a popular approach to tie together many low- end machines together as a single functional distributes system. For example, a high end machine with four I/O channels each having a throughput of 100 MB/sec will require three hours to read a 4TB data set. With Hadoop this same data set will be divided into (typically 64MB) blocks that are spread among many machines in the cluster via Hadoop Distributed File System (HDFS). It is in use at many of the world's largest online media companies including Yahoo, Facebook, Fox Interactive Media, LinkedIn and Twitter. Hadoop is entering the enterprise as evidenced by Hadoop World 2009 presentations from Booz Allen Hamilton and JP Morgan Chase. Hadoop is making its way into the federal government as well.

V. PROPOSED APPROACH

Data security means protecting data, such as a database, from destructive forces and the unwanted actions of unauthorized users often the data that is private or confidential. This data needs to be protecting from being viewed by unauthorized people. This is especially true if the data is to be sent via a public network such as the Internet. A common use case for cloud computing involves clients uploading data and computation to servers managed by third party infrastructure providers. Since the data and programs are no longer in an environment controlled by the client, private client data may be exposed to adversarial clients in the cloud server, either by accidental misconfigurations or through malicious intent. The approach used for data storage security is represented in Fig.1.

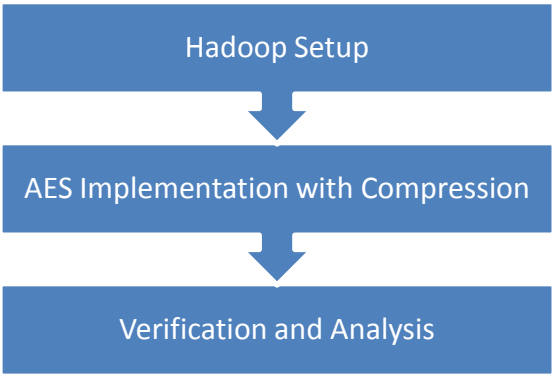

Fig.1. Data Storage Security Approach

1. Hadoop Setup

Hadoop is a framework written in Java for running applications on large cluster commodity hardware and incorporates features similar to those of the Google File System (GFS) and of the Mapreduce computing paradigm. The software versions used are Ubuntu Linux 10.04, Hadoop 1.0.3 released May 2012, and Java 1.6. Hadoop enables applications to work with huge amounts of data stored on various servers. Hadoop's functions allow the existing data to be pulled from various places and use the Mapreduce technology to push the query code and run a proper analysis, therefore returning the desired results.

2. AES Implementation with Compression

Just uploading the data is not sufficient, some modifications are required to be done when we upload the data so that data becomes unreadable and also for the purpose of security.

Compression is the reduction in size of data in order to save space or transmission time. Content compression can be as simple as removing all extra space characters inserting a simple repeat character to indicate a string of repeated characters and substituting smaller bit strings for frequently occurring characters.

Encryption technique used is symmetric encryption approach. In the proposed technique there is a common key between sender and receiver, which is known as private key. The private key concept is the symmetric key concepts where plain text is converted into encrypted text known as cipher text using private key where cipher text decrypted by same private key into plaintext.

3. Verification and Analysis

After performing the above operations i.e. compression and encryption, the results will be computed on the basis of different parameters. The most commonly used parameters will be the processing time and the space. Depending upon the parameters defined the results will be calculated for the different file formats After the development of the overall method or the model by using our proposed methodology various results will be obtained and interpret the results by using different observations.

V. IMPLEMENTATION AND RESULTS

As after the implementation of our proposed approach some of the results are obtained and these results are used for the various observations. These Observations are mentioned below. The work done by us is in Hadoop using Mapreduce Code. The key used for AES Encryption is taken as 128 bit key and the block size is taken as a 64MB by default

Observation 1: As it is known that if the file is encrypted then the size of the file may increase or decrease or may remain same so while using AES encryption the size of the file increases because there is padding at the end of the file shown in the following graph.

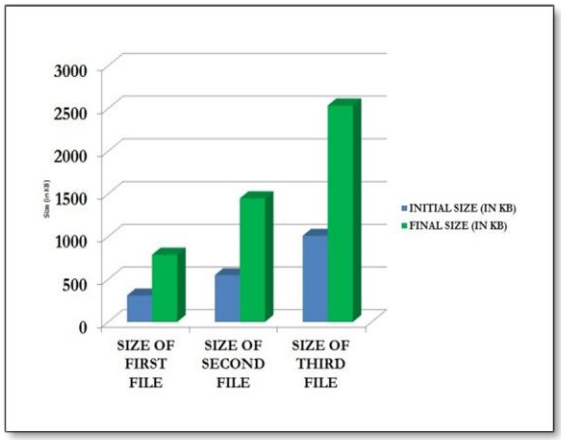

Fig.2. Comparison of sizes before and after encryption using AES

Observation 2: Fig.3 gives the performance evaluation on different datasets from which it is concluded that the time taken by the text file, audio file and the video file is almost same.

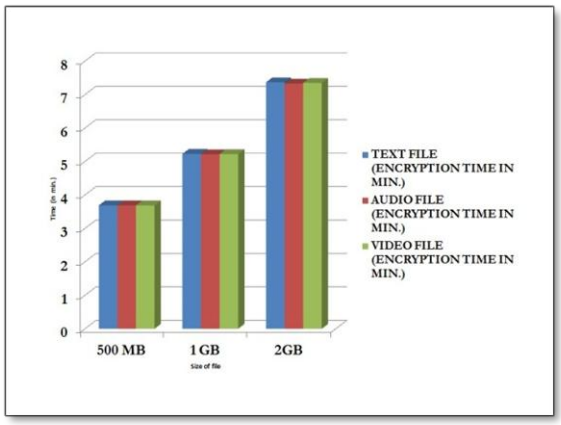

Fig.3. Performance Evaluation on Different Datasets

Observation 3:

Table 1: Results of Compression and Encryption (File size =1GB)

Type of compression	Compression before Encryption		Compression After Encryption	
	Size after Compression	Size after Encryption	Size after Encryption	Size after Compression
Gzip	338 MB	828.12 MB	2.43 GB	2.02 GB
LZ4	557 MB	2.35 GB	2.43 GB	2.35 GB

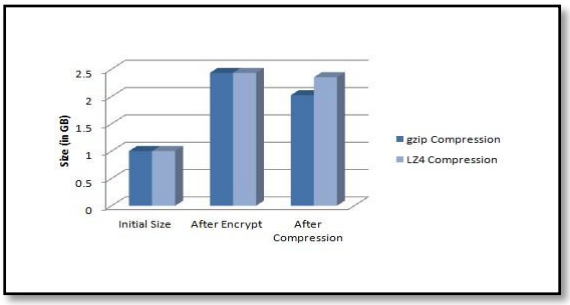

Fig.5. Representation of Encryption and Compression

The graph shown above is drawn from Table 1. It will show us the file which the client will upload that file is initially encrypted and then the file is compressed and the file compression after the encryption does not save any space while using any of the compression technique.

The following graph which is drawn from Table 1 shows us that the Gzip compression than encryption is much better than LZ4 Compression as it will save our space.

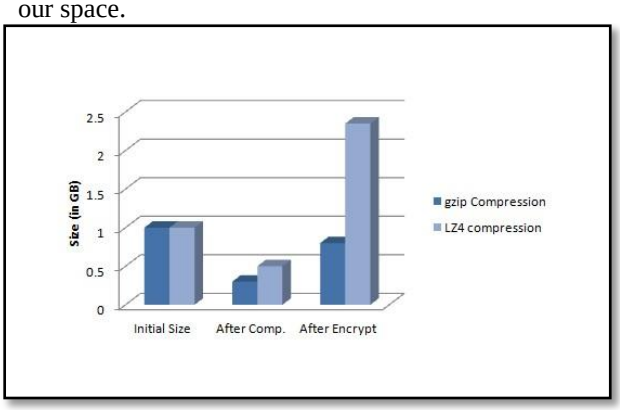

Fig.6. Representation of Compression and Encryption

VI. CONCLUSION

As protecting cloud data through encryption and to effectively utilize the resources of the cloud, compression using Mapreduce has found to be model proposed and implemented. The model is found to be adaptable to different datasets e.g. audio, video, text etc. When compared with other parallel code, Hadoop's Mapreduce is designed to parallelize user program automatically. The Hadoop's Mapreduce implementation allows performing encryption in parallel. The compression is done at the expense of time in order to utilize the resources of cloud environment. Among the various compression techniques it is concluded that Gzip compression is far better than LZ4 compression technique. Future enhancements of the work need to focus on evaluating the performance under different configuration parameters and improve a kind of encryption through selective mode. Moreover compression using other types of compression like snappy, lz0, etc. can be explored later on.

ACKNOWLEDGEMENTS

We would like to thank Dr. R.K Bawa for providing the necessary facilities for the successful completion of this study. I further thank my family and friends for their encouragement and support.

REFERENCES

[1] Anitha, P. and Palanisamy, Data protection Algorithm Using AES, *International Journal of Current Research*, 2011 *vol.3, issue 6*, pp.291-294.

[2] Chaitanya P.Sahithi and M.Murali, Improved Schemes to Secure Distributed Data Storage against Untrusted Users, *International Journal of Computer Science and Information Technology(IJCSIT)*, 2014,*vol.5(2)*, pp.1774-1777.

[3] Chauhan Nitin Singh and Ashutosh Sexena, *Cryptography and Cloud Security Challenges*, CSI Communications, 2013, pp. 18-20.

[4] Kulkarni Gurudatt, Jayanti Gambhir, Amruta Dongare, Security in Cloud Computing, *International Journal of Computer Engineering &Technology (IJCET)*, 2012, *vol. 3, issue1*, pp. 258-265.

[5] Nagendra M. and M. Chandra Sekhar , Performance Improvement of Advanced Encryption Algorithm using Parallel Computation, *International Journal of Software Engineering and its Application(IJSEIA)*,2014, *vol. 8, no 2*, pp.287-296.

[6] Purushothaman Deepanchakaravarthi, Sunitha Abburu, 2012.An Approach for data Storage Security in Cloud computing, *International Journal of computer science Issues(IJCSI), vol.9, issue 2*, no 1, pp.100-105.

[7] Rajawat Jitendra Singh, Sanjay Gaur,2013.Trusted and Secure Model for Cloud Storage, *Journal of Environment Science, Computer Science and Engineering & Technology(JECET), vol.2, no 3*, pp.883-888

[8] Rahulkar Bhavesh , Praveen Shende,2013.A Two layer encryption Approach to Secure Data Sharing in Cloud Computing, *International journal of Advanced Research in Computer Engineering and Technology(IJARCET), vol.2, issue 12*, pp. 3252-3254.

[9] Sajithabanu S., Dr. E.George, prakash Raj,Data Storage Security in Cloud, International Journal of Computer Science

and Technology (IJCST), 2011,*vol.2, issue 4*, pp.436-440.

[10] Sivashakthi T. and Dr. N Prabakaran, A Survey on Storage techniques in Cloud Computing, proceedings of *IJETAE*,2013, vol. 3, issue 12, pp.125-128.

[11] Spillner Josef, Johan Muller and Alexander Schill, 2013.Creating optimal cloud storage systems, *Future Generation Computer Systems*, pp.1062-1072.

[12] Xiao Zhifeng and Yang Xiao, Accountable Mapreduce in Cloud Computing, Proceedings of *IEEE International Workshop on Security in Computers, Networking and Communications (SCNC) in conjunction with IEEE INFOCOM*,2011, pp.1099-1104.