A TRUST BASED HYBRID ENCRYPTION APPROACH IN HADOOP

Mr. Jagdish Singh Raikwal

Department of Information Technology,

Institute of Engineering and Technology,

Indore, India

Prashul Maheshwari

Department of Information Technology,

Institute of Engineering & Technology, Indore

Abstract-Cloud computing provides some basic services (IaaS, PaaS, and SaaS) to users on the lease as per demand. Most valuable services provided by cloud computing is-it reduces hardware, maintenance and their installation cost. Users store their entire data on the cloud computing enabled storage locations and then this data gets accessed through virtual machines or mobiles or PC's. Hadoop is the biggest cloud computing and big data processing framework. It also provides the hardware where user can store their data. It is an open-source framework and increasingly used in the business and IT universe. The security is major concern in hadoop. The weakness of security mechanism in hadoop is a major concern in its development now days. The user stores its confidential data at the storage in hadoop which is a shared area and can be accessed by anyone without any security. In hadoop the data gets stored at HDFS (Hadoop Distributed File System) in the form of nodes. These data nodes can be accessed easily without any authentication and specific security. Hence the encryption of the data is a necessary thing to prevent any privacy rule violation and also to secure the confidential data from intruders.

Keywords- Hadoop, Encryption, Cloud Computing, AES, RSA, Hybrid Encryption, Hash.

1. INTRODUCTION

"The network is the computer", the generating of this thought is credited to John Gage, who was the 5th employee of Sun Microsystems. The thought really seems to be getting true. Now, everyone is putting their data on the cloud and hadoop which are not a single PC but a cluster of PCs i.e. a whole network of PCs. Hence our main concern is whether the data kept on the central storage on cloud is secure or not.

The basic aspects which we need to consider when working with data are-

Confidentiality- It says that only the sender and the recipient should be able to access the data. It gets compromised if another person is able to access the data.

Authentication- It authenticates the sender of the data. The origin of the data should be correctly identifiable. It establishes proof of identities.

Integrity- If the data, after sending by sender but before reaching to receiver, gets changed, the integrity of data is lost.

Non-Repudiation- If sender sends a message, and after that it refuses to accept that he/she had sent that data, this situation is non-repudiation.

ISSN: 2278 – 1323

Access Control- Access control ensures who should be able to see what. Based on the role and rule this setup can be done.

Availability- The principle of availability specifies that resources should be always available to authorised users.

Ethical and Legal Issues- It should deal legally with the individual's right to privacy, accuracy, property and accessibility.

2. HADOOP

Hadoop is an open source project that runs on a cluster of machines. It is open source software and its licence (Apache Open Source Licence v2.0) is very commercial friendly. It originally developed and made open source by Yahoo. Now it is developed as an Apache open source stack and contributed by many vendors like Cloudera, Hortonworks, Facebook etc.

Hadoop is known for providing both distributed storage and distributed processing of data. Hadoop storage is provided by HDFS (Hadoop Distributed File System) and the processing of data is provided by MapReduce.

HDFS- The Hadoop Distributed File System (HDFS) is the core service of Hadoop which is a distributed file system, designed to run on commodity hardware that is used in cloud. There are many similarities between HDFS and existing distributed file systems. However, HDFS has the differences from other distributed file systems which are significant and distinguish it from other distributed file systems. As HDFS is highly fault-tolerant as well as it is designed to be deployed on low-cost hardware. High throughput access to application data is provided by HDFS and that is the reason it is best suitable for large data sets applications. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. And now HDFS is an Apache Hadoop subproject. [1]

MapReduce- Hadoop uses a software framework for writing applications which process large amounts of data sets in-parallel on large number of clusters of commodity hardware in a fault-tolerant and reliable manner.

A MapReduce 'job' splits the input data set into independent chunks or blocks which are processed by the 'map tasks' in a parallel manner. The outputs of the maps are then give as input to the 'reduce tasks'. A file system is used to store the input and the output of the job. [2]

3. AES

The 'Advanced Encryption Standard (AES)' is the most used symmetric cipher technique. Advanced Encryption Standard has security strength better than 3DES and improved efficiency. In comparison to the DES which is used for legacy systems now, a new more secure environment is needed. Hence NIST called for new approaches. The requirements which NIST proposed are-

- AES should be a block cipher with 128 bit block size.
- It must support three (128, 192, 256 bit) key lengths.

The other competitors of AES were MARS, RC6, Serpent, and Two Fish. But NIST selected Rijndael as the AES.

AES is deemed secure because:

- Its building blocks and design principles are fully specified and it is efficient in both software and hardware.
- It was selected as part of an open competition.
- It has sustained 15 years of attempted cryptanalysis from many smart people, in a high-exposure situation, and it came out relatively unscathed.

ISSN: 2278 – 1323

AES has some steps to follow in encryption and decryption process-

1. Key Addition,
2. Byte Substitution, and
3. Diffusion

    3.1.1. Shift rows,

    3.1.2. Mix Column.

In encryption the steps works in the following order-

Mix column (Shift Rows (SubBytes (Text)))

However, the last round does not have the Mix Column layer which makes it symmetric in nature.

The decryption process is as follows-

SubBytes (Shift Rows (Mix Column (Cipher)))

4. RSA

RSA is public key cryptography algorithm which uses prime factorization as trapdoor one way function

$n \equiv pq$

For p and q primes also define as a private key D and public key E

$de \equiv 1 \pmod{(n)}$

Here, $a \equiv b \pmod{m}$ is congruence. Let the message be converted to a number M. The sender then makes n and e public and sends.

The RSA algorithm is totally based on the difficulty of factoring a number that is the product of two large prime numbers (almost up to 200 digits each).

However, in his paper, Rivest et al. asked the question whether it is possible to work with encrypted data, without decrypting it first. [7] This started the research for hybrid encryption system. As, RSA uses much larger keys, it takes too much time to encrypt large amount of data.

Hence a new encryption system is used which efficiently uses the RSA as it is easy to implement and best understood and also increase the efficiency of encryption and decryption in contrast to time.

5. PROPOSED APPROACH

The best way to ensure data security is 'cryptography'. Cryptography has several technologies. These techniques uses encryption method to keep the data secure from intruders. As, Hadoop is the biggest vendor of processing and storing data at large scale on cloud, the security of data is a major concern. Hence, hadoop uses some encryption techniques to ensure security. It uses AES encryption algorithm to encrypt data at rest.

As mentioned above, no any encryption algorithm provides complete security and they all have their own limitations and loopholes which cannot be ignored. Hence a new technique is introduced in this paper. This technique is based on the concept of hybridization of two or more encryption algorithms. Through this hybrid technique, the limitations of encryption algorithms are overcome and security is improved.

In this paper a hybrid approach is proposed. The encryption and decryption of data is done using that hybrid encryption algorithm.

The encryption of data has several steps as mentioned below in fig. 1-

i. First, the plain text is fed as input to the hash function which in turn generates a random key as output.
   HASH (PT) = RK
ii. This Random Key is used to encrypt the plain text using the AES encryption algorithm.
   AES (PT)

ISSN: 2278 – 1323

iii. This Random Key is then encrypted using the RSA public key algorithm.
RSA (RK)

iv. After this the encrypted key is appended to that encrypted text.
AES (PT) + RSA (RK)

v. The combined package is then again encrypted using AES which gives the final cipher text.

Final Cipher Text = AES (AES (PT) + RSA (RK))

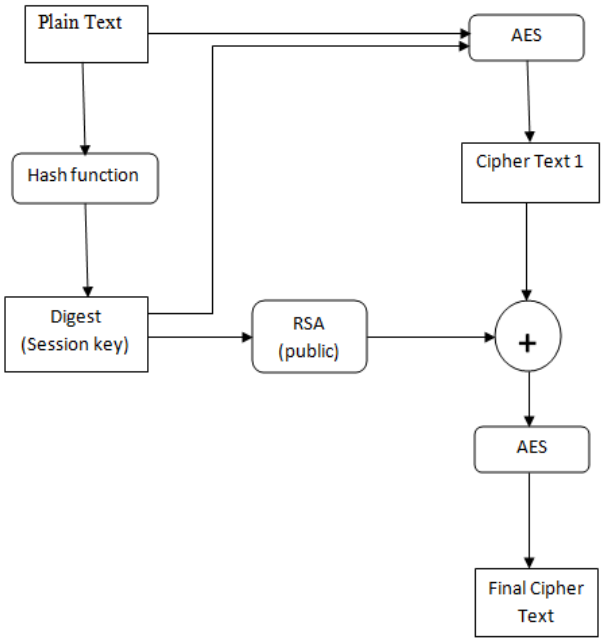

Figure 1-Encryption Process

The decryption process is just vice versa to the encryption process.

The steps are as described in fig. 2-

I. The Final Cipher Text is decrypted using AES.

II. Then the Decode process is done on the output of the first step.

III. Second step gives the Intermediate text and the RSA encrypted random key.

IV. That random key is decrypted by the RSA private key algorithm which provides the original random key.

V. That Random key is used to decrypt the intermediate text which is given by the step three using AES.

VI. The Output of the previous step is the desired Plain Text.

Plain Text = AES (AES (CT) + RSA (RK))

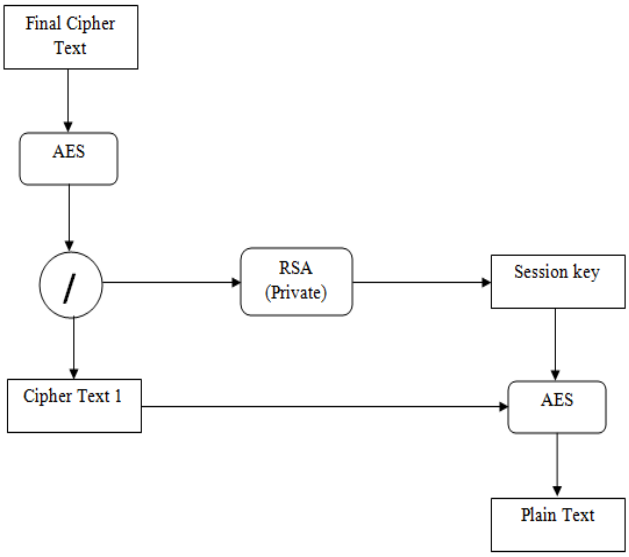

Figure 2-Decryption Process

6. CONCLUSION

Proposed algorithm is secure because it encrypts and decrypts message with one time generative key which is random in nature and gets destroyed automatically after one use. Three level of security is implemented. Algorithm is based on hybrid cryptography as it uses asymmetric and symmetric approach for both encryption and decryption processes. The AES and RSA hybrid cryptography algorithm is relatively more secure and easier to implement.

ISSN: 2278 – 1323

7. REFERENCES

1. https://hadoop.apache.org/docs/r1.2.1/hdfs_ design.html.

2. https://hadoop.apache.org/docs/current/had had-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html.

3. Charanjeet Kaur et al, "Hybrid Encryption Scheme for Hadoop Based Cloud Data Security", IJCSET, July 2015, Volume 5, Issue 7, 250-254.

4. Ritu Pahal, "Efficient Implementation of AES", IJARCSE, Volume 3, Issue 7, 2013.

5. Shaomin Zhang, Yufang Gan, Baoyi Wang, "Parallel Optimization the AES Algorithm Based on MapReduce", scientific, 644-650, 2014.

6. http://www.drdobbs.com/security/the-hmac-algorithm/184410908.

7. R. Rivest, A. Shamir, L. Adleman, "A Method for Obtaining Digital Signature and Public-Key Cryptosystems", 1978, Communications of the ACM, Volume 21.

8. R. Rivest, A. Shamir and L. Adleman, "A Method for Obtaining Digital Signatures and Public Key Cryptosystems", 1999, Communications of the ACM, Computer Science, pages 223-238, Springer.

ISSN: 2278 – 1323