DATA SECURITY ON HADOOP DISTRIBUTED FILE SYSTEM USING RSA ENCRYPTION

Mr. A. Antony Prakash[1*], Dr. A. Aloysius [2]
[1] Department of Information Technology, St.Joseph's college, Trichy,India
[2] Department of Computer Science, St.Joseph's college,Trichy, India

Abstract : The Public key cryptography is becoming a trending revolution in the area of cryptography for the authentication and maintaining confidentiality of the information. When comparing various cryptosystem RSA method of key generation proves to be more efficient with respect to time and security. The Hadoop Distributed File System (HDFS) is the stockpiling framework required by Hadoop applications. Huge information was analysed, supported by hadoop.The proposed system employs innovative encryption decryption techniques with the use of RSA for key generation and Hadoop for data storage. The tested method was compared with few prevailing techniques with respect to computational time for key generation and time required for encryption and decryption and proved that our proposed system consumes less time than the existing.

IndexTerms **: HDFS RSA Encryption, Cipher text, Plain text, and symmetric key**

I. INTRODUCTION

Cryptosystem is a traditional and fine technique for the purpose of information hiding and data security[1].In recent years, the public key generation by RSA is extensively utilized to support the storage and processing of the data and the management of the information in the Internet of Things (IoT).

Hadoop system performance was improved in reliable manner. Hadoop stores huge amount of data. Complex dataset was processed to increased system speed, performance. Hadoop was introduced by apache foundation. Hadoop works in reliable manner. Security of data was retained by Hadoop distributed file system by RSA Algorithm. Rivest- Shamir-Adleman is algorithm for encryption, decryption of messages. RSA algorithm maintains security measures such as Authentication, confidentiality, Authorization. Basically, cryptographic algorithm are low cost, high performance[2].

Time, processor, memory are required less for RSA algorithm. Calculation involved in RSA algorithm is simple. Time consumption, resources are less along with less complex structures was the greatest advantages in RSA algoirthm.RSA algorithm is asymmetric key encryption. It is a cryptosystem form for doing encryption, decryption using different keys such as one public key, one private key. Asymmetric key encryption was also called as Public key encryption. Ron rivest, Adi shamer, Leonard adleman, was introduced RSA encryption algorithm and named after them. Some of the RSA generated key does not uses the rejection sampling, the conventional algorithm which chooses a random number. RSA algorithm includes three stages a) Key generation b) encryption c) Decryption. Encryption is weak when small p and q values occurs. Time consumption is lot with decreased performance, when large p, q values occurs. The proposed work deals with the key generation via RSA algorithm and storage of the encrypted data in Hadoop thereby improving the data security.

Objective:

- To store the encrypted data into Hadoop File System.
- To confirm the file security through key generation by RSA algorithm.
- To make easy, secure and authenticated access for the users who need the data.

1.1 organization of the Paper

The remaining division is specified as mentioned below: Section II provides the brief explanation of the existing methods relevant to the research work is discussed in the related works. Section III deliberates the implementation process of encryption and decryption techniques of the proposed system Section IV illustrates the comparative investigation of projected methodology with the prevailing methods and accomplishes the research work based on our results and examination. Section V concludes this research work

II RELATED WORKS

This portion provides a discussion about existing works and their merits and demerits. From the examination of the existing works it is observed that the RSA and Hadoop could be readily employed for the data security study.
[3]investigated the fundamental knowledge regarding the RSA based algorithm and compared the existing techniques on the basis of few factors like attack vulnerability, technique uniqueness and so on. It is observed that this RSA algorithm has more security with the implementation of the few techniques. In in paper it is recommended to add the image pixel to make the algorithm to be more powerful.

[4] proposed a comparative study in between the RSA cryptosystem that is a asymmetric cryptosystem and the two symmetric system named blowfish and Test.The outcomes showed that the duration of the time taken by the use of analytical relations in the RSA and enabling speedy application than the Blowfish and DES with highly secured information. However in RSA the attribute of the prime

number chosen (Q and P) regulates the time in the process of key generation thereby increasing the time taken for the security purpose than before.

[5]Investigated an enhanced and modified model on the basis of RSA public key cryptosytem.The work utilizes four large prime numbers that adds the system complexity when compared with the conventional model. Here n is the two large prime number product and the four prime numbers provides the value for the encryption and decryption which improves the security of the system. The suggested system required more time for cryptographic analysis which is more than the conventional RSA algorithm. The comparative study proved to more efficient.

[6] illustrated data storing with processing huge data quantity was done by open source named hadoop. File encryption was done by hybrid encryption scheme where files are symmetrically encrypted with image secret key followed by asymmetric key encryption by public key. Private Key was retained by user, encrypted file was sent to HDFS. In this File Encryption and Data Key the chance of image characters is verified before key generation usage. In File Decryption and Data Key Acquisition. While downloading file by requesting user, File from Hadoop distributed file system was sent to server by api calls, data key was obtained by requesting response data key management module. Data key was restored by private key and used for file decryption .Symmetric encryption key algorithm increases performance, data security by hadoop dependent cloud data security.

[7] presented function test, extensibility test and efficiency test verified by distributed encryption algorithm. In distributed decryption algorithm, Task was performed by server by encrypted fragment, decryption algorithm. Decryption of every task was done by RSA. The cipher text fragment was deciphered by decryption algorithm obtains corresponding plaintext fragment. Extracted fragments was detected depending own cipher text fragments, grouping of identities in clear text fragments, cipher text fragments. Encryption size varies with plain text size. Map reduce is merged with RSA encryption algorithm forms Distributed encryption in RSA algorithm.

[8] introduced the executing security in big data by hadoop server. User Interface view explains about the graphical interface of supermarket application with source code module. Action and description with used packages and classes information was mentioned in user interface view. The front end application includes login page, registration page, and the operational interface. Users are given credentials, access by control mechanism. It comprises of java file with its associated packages. Privacy, credentials transparency was maintained in data distribution by logical view. Back ground view desires are achieved by completed execution. User rights, responsibility was carried out in logical view. User operation for supermarket was done by background view.

[9] developed an investigation on the hyper spectral images acquired by remote sensing satellites was input to proposed system. The processing of input images was performed by using Orfeo toolbox for producing Geo-Spatial Database. Followed by that normalization of geo spatial database, MapReduce and Hadoop to accumulate it in cloud. Finally, processing of the Map Reduced Geospatial data in different applications was done. Cloud data was accessed by simple authentication .The output was represented in either raster or in vector form depending upon application query provided to client.

[10] represented the securing big data by RSA, AES, DES algorithm. Deployment cloud computing Platform as a Service (PaaS), Infrastructure as a Service (IaaS), platform for customers to develop, run and manage applications in absence of owning the infra was provided by Platform as a service. Business applications included in cloud virtual servers was done by SAAS. There are two types of node in hadoop namely name node, data node. Data distribution in data node was performed by name node. If the user requests data, the encrypted data is sent to the server for decryption. Encrypted, decrypted data with user keys was provided to user.

[11] presented we are proposing a technique of introducing Validation Lamina in Hadoop system that will review electronic signatures from an access control list of concerned authorities while sending & receiving confidential data in organization. If Validation gets failed, concerned authorities would be urgently humiliated by the system and the request shall be repeatedly put on halt till required action is not taken for privacy governance by the authorities. Authentication: It is analogous to signing a contract but in digital world. Authorization: It is analogous to signing & identifying one's own signature in digital world. Dark data or computed data which is accessed by user possess almost same level of technical configurations in Hadoop environment as non-Hadoop implementations.

[12] illustrated two main components of Hadoop: HDFS and Map Reduce. HDFS Client provides interface between user and Hadoop. HDFS Client communicates with name node (via heartbeat messages) name node finds appropriate data node, name node provides details of data node, HDFS Client upload file to data node that divides files into blocks and stores it. It makes three Replicas of that file which are data node provides blocks details to Name Node. the main issue to be addressed is how to identify duplications and how to prevent duplicates from uploading to HDFS. SHA algorithm was used to mark a single impression for all file and fast fingerprint index setup in HBase the duplications are identified. Hadoop database (HBase) which was an distributed, versioned and column-oriented database open-source. The frequent large scale engineering applications uses HDFS. HDFS and HBase are the features that act as a storage and indexing system used in our work.

III. EXISTING SYSTEM

There are numerous methods available for data security and data storage. The conventional cryptographic methodologies and the existing algorithms for examined with respect to their key size and computation time[13]. Blowfish and AES algorithm has been chosen as existing in the work. These algorithm provide better efficiency and high security with respect to intrusion. The merit of the existing algorithm is that the key size is longer which offers security, however the demerit is that the existing system is very slow for the process of encryption an ecryption.Our proposed work attempts to overcome the reduced time problem with the existing system. The encryption and decryption values of the proposed and existing[14]were compared and the results were shown in the performance analysis.

[13, 15]In this existing system, DAC-MACS and DAC system are investigated that overcomes two kinds of attack: The first is that interruption with the updating of user keys thereby obtaining the proper token key to decrypt the secret data and the second is that

interception o cipher text update key to recover the capability to decrypt the stores secret information. However this process also lacks in decryption time the performance analysis of the existing was compared with the proposed and proved to be more efficient with respect to time.

IV PROPOSED SYSTEM

In our Proposed Work, encryption and decryption are done with TRSA algorithm. RSA algorithm is asymmetric cryptography algorithm works with two diverse keys such as Private Key, Public Key. From the name itself it is known that Public Key is provided to all and the private Key is retained as private. Plain text data conversion (plaintext) into random, meaningless appearance (cipher text) was known as Encryption. Cipher text was converted back to plaintext by decryption. The aim of proposed system is resolve the difficulty in decryption of generated cipher text in absence of key. Encryption is vital because data security and accessing information from unauthorized access was restricted, confidentiality and thus maintains the confidentiality.

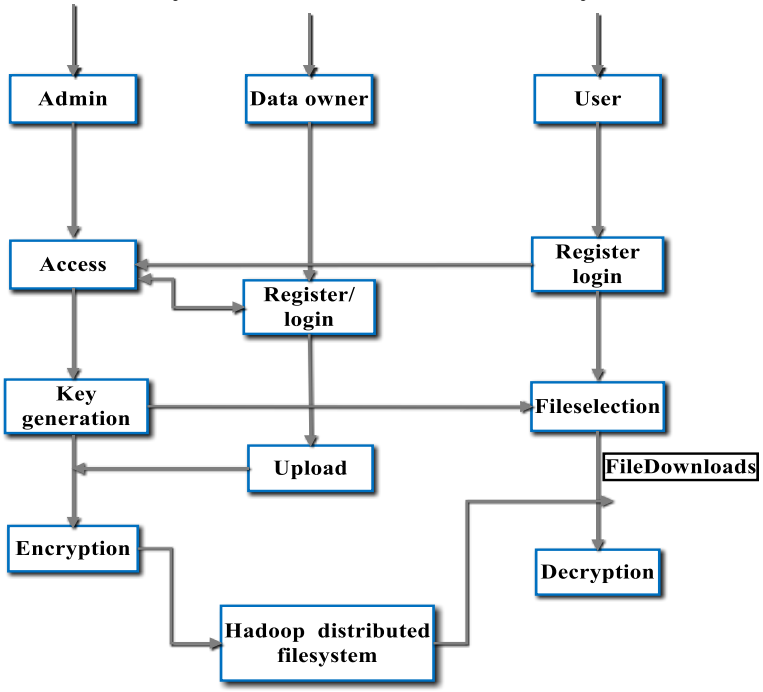


Fig 1: Flow chart depicting the proposed system

V PERFORMANCE ANALYSIS

The performance of considered technology are discussed in this section.

Performance Analyzer:

- Encryption Time
- Decryption Time

INPUT: Required modulus bit length, rr.

OUTPUT: An RSA key
RSA algorithm is asymmetric cryptography algorithm. Works with two different keys i.e. **Public Key** and **Private Key.** From the name itself it shows Public Key is provided to everyone and Private Key is maintained as private.

The RSA algorithm Key Generation

Step 1: INPUT: Selecting two large prime numbers t and s.
Step 2: Calculate the system modulus, w = s*t and the 'totient' function $\varphi(w)=(s -1)(t-1)$. It should be noted that the factors s and q persist secret and w is public.
Step 3: Choose the encryption key ek arbitrarily, so that $\gcd(ek,\varphi(w))=1$, where $1< ek <\varphi(n)$.
Step 4: Resolve the resulting equation to calculate the decryption key dt: $e.d =1 \bmod \varphi(w)$, where $0 \le dt \le w$.
Step 5: Publish the public encryption key: PUKY = {ek, w}that is depicted to all.
Step 6: Hold private or secret the decryption key: PR = {dt, w}, that is well-known only to the individual who has to sign the message or decrypt.

Data Encryption

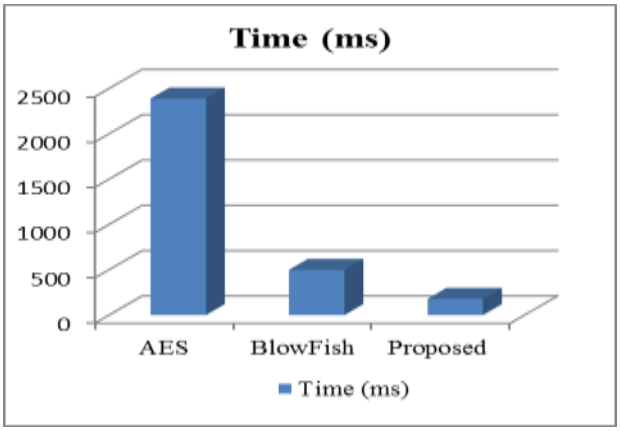
Step 1: Ms← Message
Step 2: W← Total count Message
Step 3: C← Cipher Text
Step 4: Input the plaintext or message Ms, where 0 ≤ Ms ≤ w .
Step 5: Obtain the public key of recipient, PU = {en, w}.
Step 6: Compute the cipher C, using the following equation: C = Ms en mod w

Data Decryption
Step 1: C ← Cipher Text
Step 2: PRi← Private Key
Step 3: M← Message
Step 4: Input the cipher text C.
Step 5: Use their private key, PRi = {en, w}.
Step 6: Compute the message M, using the following equation: M = C^{en} d mod w
AES = Advance Encrypte Standard

Table 1: RSA Encryption Time

Algorithm	AES	BlowFish	Proposed
Time (ms)	2390	500	180

Fig 2: Encryption Time

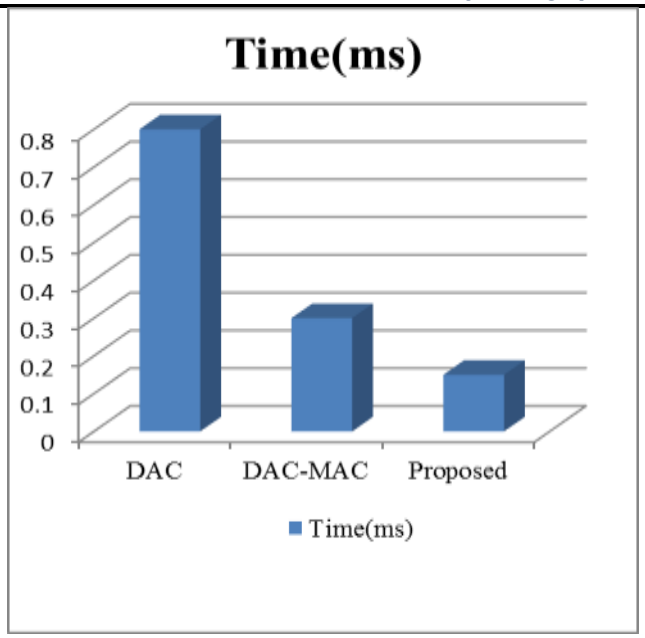


The figure 2 explains that the proposed framework requires less encryption time than the existing AES and BlowFish technique.

Table: 2 Data Decryption time

Algorithm	**DAC**	**DAC-MAC**	**Proposed**
Time(MS)	0.8	0.3	0.15

Fig 3: Decryption Time



The figure 3 explains that the proposed framework requires less decryption time than the existing AES and BlowFish technique.

VI CONCLUSION

By implementing proposed systems we achieve great encryption decryption methodology using RSA. The proposed system was compared with some existing methodology by comparing the computation time for key generator, encryption and decryption time. Huge information with delicate and private data, was ensured in different product equipment by verification, client confirmation. Information was collected from different source in big data. Hadoop was used in various ventures for information processing, security arrangement. In this manner validation, approval and encryption or unscrambling techniques are much accommodating to verify Hadoop record framework.

REFERENCES

[1] Z. Yan, M. Wang, Y. Li, and A. V. Vasilakos, "Encrypted data management with deduplication in cloud computing," *IEEE Cloud Computing*, vol. 3, pp. 28-35, 2016.

[2] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of Big Data*, vol. 2, p. 24, 2015.

[3] S. Khatarkar and R. Kamble, "A survey and performance analysis of various RSA based encryption techniques," *International Journal of Computer Applications*, vol. 114, 2015.

[4] A. El-Deen, E. El-Badawy, and S. Gobran, "Digital image encryption based on RSA algorithm," *J. Electron. Commun. Eng*, vol. 9, pp. 69-73, 2014.

[5] M. Thangavel, P. Varalakshmi, M. Murrali, and K. Nithya, "An enhanced and secured RSA key generation scheme (ESRKGS)," *Journal of information security and applications*, vol. 20, pp. 3-10, 2015.

[6] D. Shehzad, Z. Khan, H. Dağ, and Z. Bozkuş, "A novel hybrid encryption scheme to ensure Hadoop based cloud data security," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 14, 2016.

[7] Y. Xu, S. Wu, M. Wang, and Y. Zou, "Design and implementation of distributed RSA algorithm based on Hadoop," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-7, 2018.

[8] S. Singh and M. Sharma, "The Prototype for Implementation of Security Issue in Big Data Application using Hadoop Server," *International Journal of Computer Applications*, vol. 145, 2016.

[9] K. Sankar and P. Sevugan, "An investigation on hybrid computing for competent data storage and secure access for geo-spatial applications," *IIOAB journal*, vol. 7, pp. 139-149, 2016.

[10] I. Bhargavi, D. Veeraiah, and T. M. Padmaja, "Securing BIG DATA: A Comparative Study Across RSA, AES, DES, EC and ECDH," in *Computer Communication, Networking and Internet Security*, ed: Springer, 2017, pp. 355-362.

[11] R. Kumar, D.-N. Le, and J. M. Chatterjee, "Validation Lamina for Maintaining Confidentiality within the Hadoop," *International Journal of Information Engineering and Electronic Business*, vol. 11, p. 42, 2018.

[12] P. Prajapati, P. Shah, A. Ganatra, and S. Patel, "Efficient Cross User Client Side Data Deduplication in Hadoop," *JCP*, vol. 12, pp. 362-370, 2017.

[13] G. Ragesh and K. Baskaran, "Cryptographically Enforced Data Access Control in Personal Health Record Systems," *Procedia Technology*, vol. 25, pp. 473-480, 2016.

[14] D. S. A. Rakhi Emelaya "A Survey: Secure Data Storage Techniques in Cloud Computing " *International Journal on Recent and Innovation Trends in Computing and Communication* vol. Volume: 3 2015.

[15] X. Wu, R. Jiang, and B. Bhargava, "On the security of data access control for multi authority cloud storage systems," *IEEE Transactions on Services Computing*, vol. 10, pp. 258-272, 2015.