International Journal of Advanced Research in Engineering Technology & Sciences ISSN: 2394-2819
June-2016 Volume 3, Issue-6
Email:editor@ijarets.org www.ijarets.org

A HYBRID APPROACH FOR SECURITY IN BIG DATA USING RSA AND MODIFIED AES

Vandana Sehgal
M.Tech
Computer Science Department
Ganga Institute of Technology and Management
Kablana ,Jhajjar, Haryana
Maharashi Dayanand University
Rohtak, Haryana

Dr. Yashpal Singh
Assistant Professor
Computer Science Department
Ganga Institute of Technology and Management
Kablana ,Jhajjar, Haryana
Maharashi Dayanand University
Rohtak, Haryana

Abstract: **-**Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. This research is particularly valuable for inexperienced solution providers like universities and research organizations, and will allow them to swiftly set up their own big data storage services. The time consumption will vary in both cases as input file is small and large in size. So there are two comparisons one is between RSA and Modify AES cipher and other between both sizes of file.

Keyword: - Cryptography, RSA, AES, Big Data, Networking

I. INTRODUCTION TO CRYPTOGRAPHY

The fundamental and classical task of cryptography is to provide confidentiality by encryption methods. The message to be transmitted it can be some text, numerical data, an executable program or any other kind of information is called the plaintext. Alice encrypts the plaintext m and obtains the ciphertext c. The ciphertext c is transmitted to Bob. Bob turns the ciphertext back into the plaintext by decryption. To decrypt, Bob needs some secret information, a secret decryption key.1 Adversary Eve still may intercept the ciphertext. However, the encryption should guarantee secrecy and prevent her from deriving any information about the plaintext from the observed ciphertext.

Encryption is very old. For example, Caesar's shift cipher2 was introduced more than 2000 years ago. Every encryption method provides an encryption algorithm E and a decryption algorithm D. In classical encryption schemes, both algorithms depend on the same secret key k. This key k is used for both encryption and decryption. These encryption methods are therefore called

1. Sometimes the terms encipher and decipher are used instead of encrypt and decrypt.
2. Each plaintext character is replaced by the character 3 to the right modulo 26, i.e., a is replaced by d, b by e,. . ., x by a, y by b and z by c.

II. LITRETURE SURVEY

In 2010, enterprises and users stored more than 13 Exabyte of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [1]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D [2] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [3], the New York Times [4], and National Public Radio [5, 6].
While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [7], and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modeling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase. For example, Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model. Doing so raises the need to track provenance and to handle uncertainty and error: topics that are crucial to success, and yet rarely mentioned in the same breath as Big Data. Similarly, the questions to the data analysis pipeline will typically not all are laid out in advance. We may need to figure out good questions based on the data. Doing this will require smarter systems and also better support for user interaction with the analysis pipeline. In fact, we currently have a major bottleneck in the number of people empowered to ask questions of the data and analyze it [4]. We can drastically increase this number by supporting many levels of engagement with the data, not all requiring deep database expertise. Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how we manage data analysis.

III. RSA ALGORITHM

RSA is one of the first practicable public-key cryptosystems and is widely used for secure data transmission. In such a cryptosystem, the encryption key is public and differs from the decryption key which is kept secret. In RSA, this asymmetry is based on the practical difficulty of factoring the product of two large prime numbers, the factoring problem. RSA stands for Ron Rivest, Adi Shamir and Leonard Adleman, who first publicly described the algorithm in 1977. Clifford Cocks, an English mathematician, had developed an equivalent system in 1973, but it wasn't declassified until 1997.
A user of RSA creates and then publishes the product of two large prime numbers, along with an auxiliary value, as their public key. The prime factors must be kept secret. Anyone can use the public key to encrypt a message, but with currently published methods, if the public key is large enough, only someone with knowledge of the prime factors can feasibly decode the message. Breaking RSA encryption is known as the RSA problem. It is an open question whether it is as hard as the factoring problem.

IV. AES

Advanced Encryption Standard, is a symmetric block cipher that can encrypt data blocks of 128 bits using symmetric keys 128, 192, or 256. AES encrypt the data blocks of 128 bits in 10, 12 and 14 round depending on the key size. Brute force attack is the only effective attack known against this algorithm. AES encryption is fast and flexible.

AES is the new encryption standard recommended by NIST to replace DES in 2001. AES algorithm can support any combination of data (128 bits) and key length of 128, 192, and 256 bits. The algorithm is referred to as AES-128, AES-192, or AES-256, depending on the key length. During encryption-decryption process, AES system goes through 10 rounds for I28-bit keys, 12 rounds for I92-bit keys, and 14 rounds for 256-bit keys in order to deliver final cipher-text or to retrieve the original plain-text [8]. AES allows a 128 bit data length that can be divided into four basic operational blocks. These blocks are treated as array of bytes and organized as a matrix of the order of 4×4 that is called the state. For both encryption and decryption, the cipher begins with an Add Round Key stage. However, before reaching the final round, this output goes through nine main rounds, during each of those rounds four transformations are performed 1) Sub-bytes, 2) Shift-rows, 3) Mix-columns, 4) Add round Key. In the final (10th) round, there is no Mix-column transformation [9], [10]. Fig. 3 shows the over-all process. Decryption is the reverse process of encryption and using inverse functions: Inverse Substitute Bytes, Inverse Shift Rows and Inverse Mix Columns. Each round of AES is governed by the following transformations [11].

1. **Substitute Byte Transformation**

AES contains 128 bit data block, which means each of the data blocks has 16 bytes. In sub-byte transformation, each byte (8-bit) of a data block is transformed into another block using an 8-bit substitution box which is known as RijndaelSbox.

2. **Shift Rows Transformation**

It is a simple byte transposition, the bytes in the last three rows of the state, depending upon the row location, is cyclically shifted. For 2nd row, 1 byte circular left shift is performed. For the 3rd and 4th row 2-byte and 3-byte left circular left shifts are performed respectively.

3. **Mix Columns Transformation**

This round is equivalent to a matrix multiplication of each Column of the states. A fix matrix is multiplied to each column vector. In this operation the bytes are taken as polynomials rather than numbers.

4. **Add Round Key Transformation**

It is a bitwise XOR between the 128 bits of present state and 128 bits of the round key. This transformation is its own in-verse.

V. OBJECTIVE OF RESEARCH WORK

Step 1: It is necessary to study the concept of cryptography in Big Data. How a big data challenge security and how can we face this challenge?

Step 2: Encryption process is applied on a file so we take two file whose volume varies but variety and velocity remain constant.

Step 3: RSA algorithm is applied on those files one by one and time consumption in both cases is compared. This comparison finds out effect of volume variation on encryption process.

Step 4: Now AES is modified and applied on those files one by one and time consumption in both cases is compared. This comparison finds out effect of volume variation on encryption process.

Step 5: A table generated in which comparison for RSA and Caesar cipher is shown for both files.

VI. FLOWCHART FOR PROPOSED METHOD

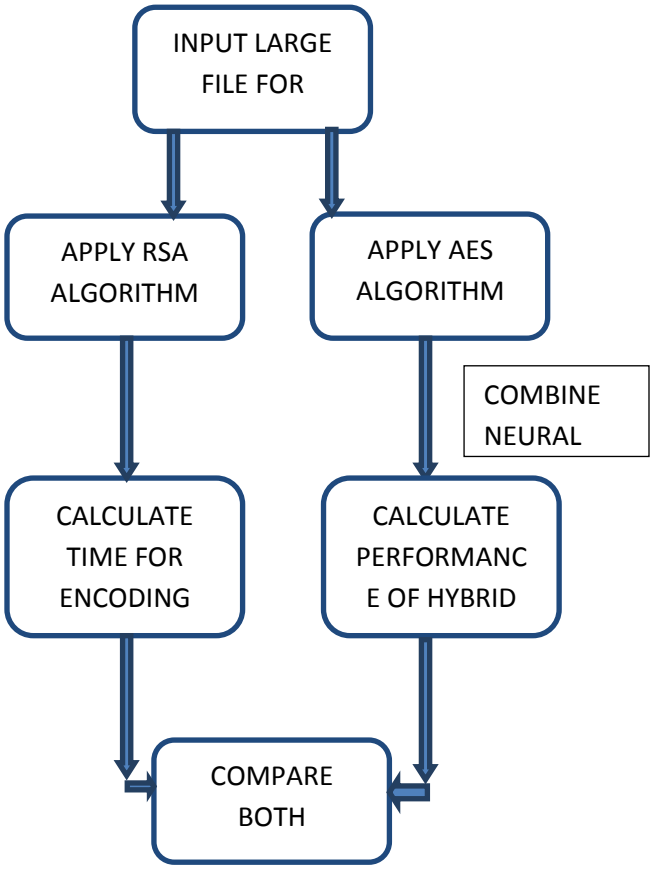
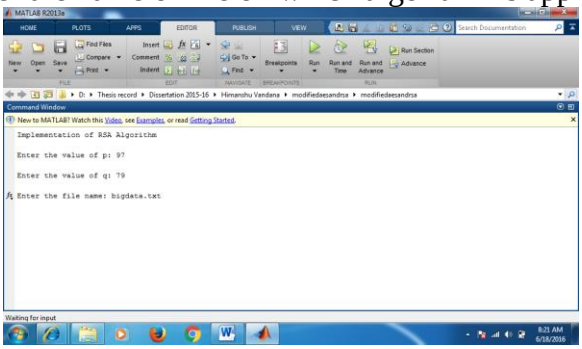


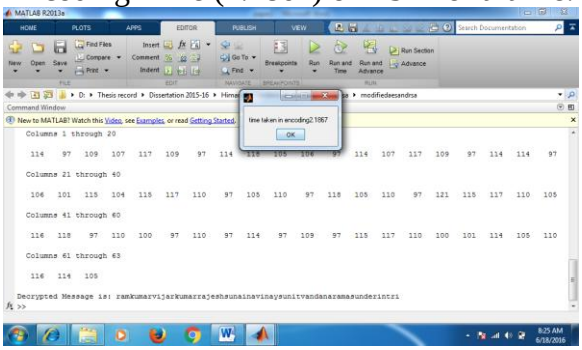
Fig 1: Steps follow to achieve proposed result

VII. RESULTS

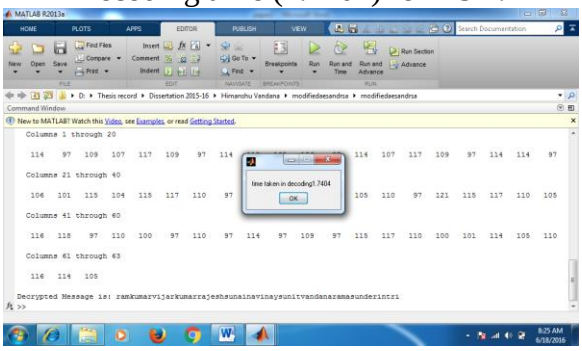
Enter the name of file on which algorithm is applied:



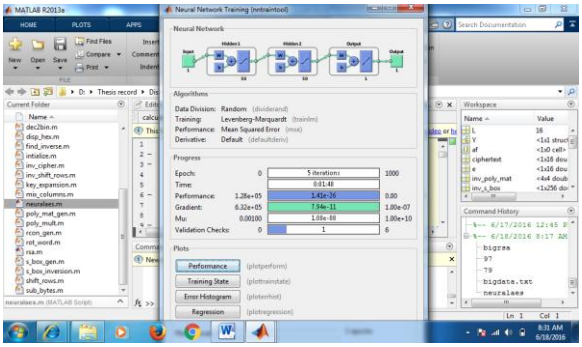
Encoding Time (2.1867) of RSA for txt file:



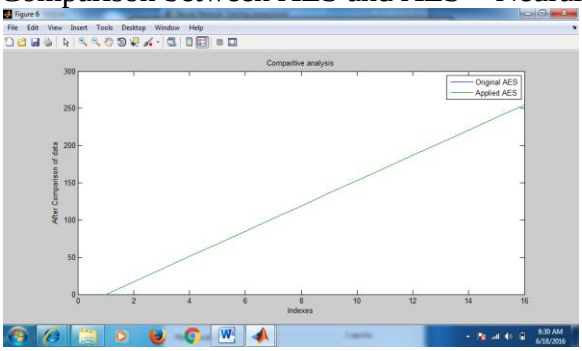
Decoding time (1.7404) for RSA:



No of iterations are 5 in AES + Neural



Comparison between AES and AES + Neural



VIII. CONCLUSION & FUTURE SCOPE

Big data privacy has become an important issue since it is directly related to customers. It is now essential for an organization to promise privacy in big data analytics. Privacy measures should now focus on the uses of data rather than collection of data. They should be modified with respect to the size and unexpected uses of big data. Techniques like anonymization have limited potential when applied to big data. Notice and consent method also burdens the customer for ensuring privacy. Differential privacy may be seen as a viable solution for big data privacy. One problem with this method is that analyst should know the query before using the differential privacy model. When modified and applied to big data, it may ensure privacy without actually modifying the data.

REFERANCE

1. Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, McKinsey Global Institute, May 2011.
2. Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. PCAST Report, Dec. 2010.
3. Drowning in numbers -- Digital data will flood the planet—and help us understand it better. The Economist, Nov 18, 2011.
4. The Age of Big Data. Steve Lohr. New York Times, Feb 11, 2012.
5. Following the Breadcrumbs to Big Data Gold. Yuki Noguchi. National Public Radio, Nov. 29, 2011.
6. The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. National Public Radio, Nov. 30, 2011.
7. Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release, July 2011.
8. Mr. Gurjeevan Singh, Mr. Ashwani Singla and Mr. K S Sandha, "Cryptography Algorithm Comparison for Security Enhancement in Wireless Intrusion Detection System", International Journal of Multi-disciplinary Research, Vol.1 Issue 4, pp. 143-151, August 2011.
9. William Stallings, "Cryptography and Network Security: Principles and Practice", Pearson Education/Prentice Hall, 5th Edition.
10. Zilhaz Jalal Chowdhury, DavarPishva and G. G. D. Nishantha, "AES and Confidentiality from the Inside Out", the 12th International Con-ference on Advanced Communication Technology (ICACT), pp. 1587-1591, 2010
11. Akash Kumar Mandal, Chandra Parakash and Mrs. ArchanaTiwari, "Performance Evaluation of Cryptographic Algorithms: DES and AES", IEEE Students' Conference on Electrical, Electronics and Com-puter Science, pp. 1-5, 2012.