RESEARCHERID
THOMSON REUTERS
[COTII 2019]
ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

ENCRYPTION ALGORITHM FOR BIG DATA SECURITY

Shweta Sinha[1] & Priya Verma[2]
[1&2]National P.G. College

ABSTRACT

Big data is a driver of the world economic and societal changes. The world's data collection is reaching a critical point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. While the data complexities are increasing including data's volume, variety, and velocity, the real impact hinges on our ability to uncover the `value' in the data through Big Data Analytics technologies. In this paper, we explain about the big data, its evolution and the results obtained after applying a systematic mapping study to security in the Big Data ecosystem.

Keywords: *Output Transformation, Sub-Key, Symmetric Key Algorithm, round, encryption.*

I. INTRODUCTION

Evolution of data

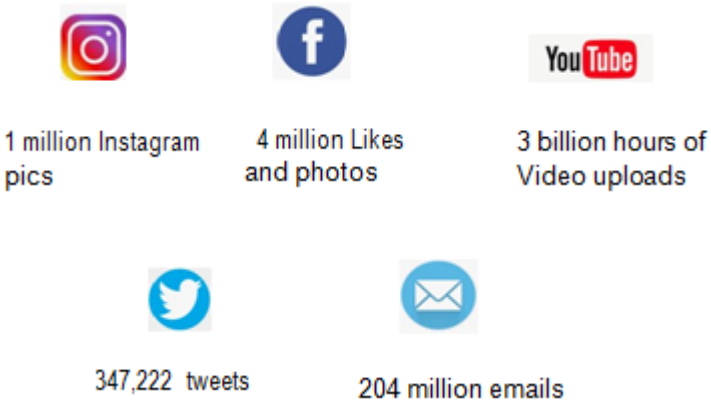
- Evolution oftechnology
- Social media
- IOT
- Other factors

Evolution of Technology:

Earlier we used landline telephones for communication and now we are using smart phones making our life smarter. Every action performed by us on mobile phones generates data. We were also using bulky desktops for processing MBs of data. We were using floppy disks to store 1-2 MBs of data and then hard disks to store TBs of data and now we are storing the data on clouds. ***Cloud Storage*** *is a service where data is remotely accessed, maintained, managed, and Backed up. The service allows the users to store files online, so that they can access them from any location using the Internet.* Today even self-driving car has come up. These cars have sensors through which they can sense the size of the obstacles, distance of obstacle and many more and then it decides how to react which again generates a lot of data each kilometer.

ISSN 2348 – 8034
Impact Factor- 5.070

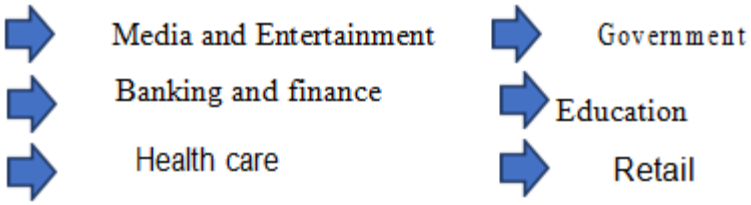
Social Media:



IOT:

IOT connects the device with internet and makes the device smarter. Today we have smart ACs, smart cars, smart phones, smart meter and many more .If we talk about smart ACs, it monitors our body temperature and accordingly decides what should be the temperature of the room. Now in order to do this it has to accumulate data either from internet or through sensors that are monitoring our body temperature. So, using IOTs we are generating a lot of data.

Other factors:



II. WHAT IS BIG DATA?

Big data is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database tools or traditional data processing applications.

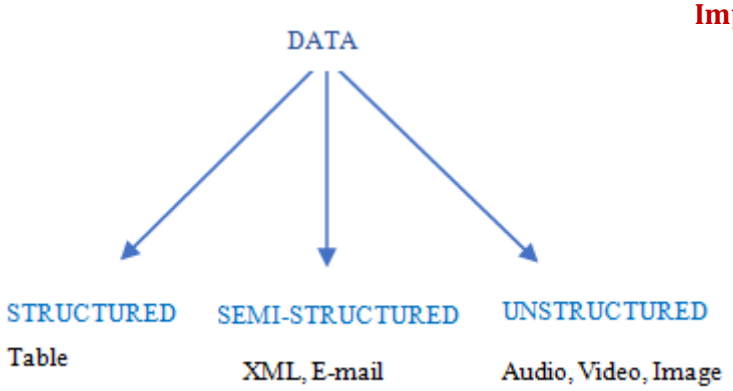
How do we consider data as big data?

There are 5 Vs to determine the same[2]:

Volume: The data is rising exponentially. Today we are dealing approximately 21000 Exabytes of data.

Variety: Different types of data are being generated from various sources.

ISSN 2348 – 8034
Impact Factor- 5.070



- Structured: Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made accessed for more effective processing and **analysis**. In this format, we have a proper schema defined for the data.
- Semi-structured: Semi-structured data is data that has not been organized into a specialized repository, such as a database, but yet has associated information, such as metadata(data about data), that makes it more easy to process than raw data.
- Unstructured: Unstructured data (or unstructured information) is information that does not have a pre-defined schema. Example- a word document, until and unless metadata tags are not added.

Velocity: Data is being generated at an alarming rate every minute.

Value: When we talk about value, we're referring to the worth of the data being extracted. Having untold amounts of data is one thing, but unless it can be turned into value it is useless. The most important part of starting on a big data initiative is to understand the costs and benefits of collecting and analyzing the data to ensure that ultimately the data that is procured can be monetized.

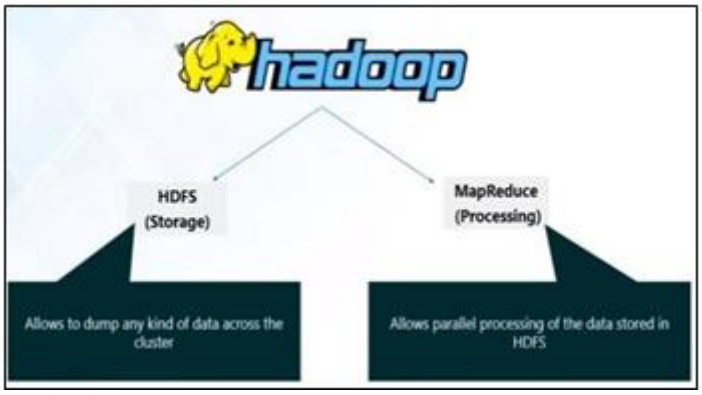
Veracity: Veracity is the quality or faithfulness of the data. A good example of this relates to the use of GPS data. Often the GPS will "shift" off course as we go through an urban area. Satellite signals are lost as they bounce off tall and large buildings or other structures. When this happens, location data has to be fused with another data source like road data, to provide accurate data.

Problems with big data:

1. Storing exponentially growing huge datasets.
2. Processing data having complex structure
3. Processing data faster

One of the solutions to the above-mentioned problems is open source software Hadoop [6]. Apache Hadoop is a collection of open source software utilities that facilitate using a network of several computers to solve problems involving huge amounts of data and computation. It provides a framework for distributed storage and processing of big data using the Map Reduce programming model [7].

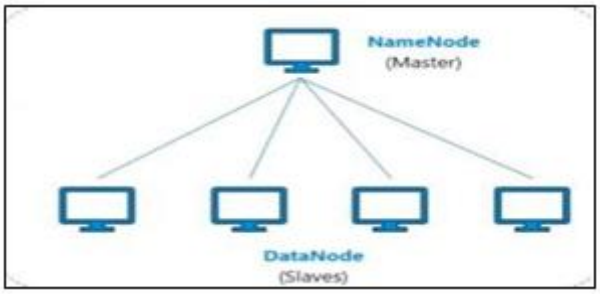
ISSN 2348 – 8034
Impact Factor- 5.070



HDFS (Hadoop Distributed File System) has two core components:

- The Name Node is main node which contains metadata about the data stored.

Data is stored on the Data Nodes which are commodity hardware in the distributed environment.



The first problem i.e. **Storing exponentially growing huge datasets** is solved using Hadoop as it divides a large file into small chunks and then stores them on different data nodes and these data files are also replicated so as to make the accessing of data easier. More no. of data nodes can be added as per the requirement.

The second problem i.e. **processing data having complex structure (storing unstructured data)** is solved as HDFS allows storing any kind of data be it structured, semi-structured or unstructured.

The third problem i.e. **Processing data faster** is solved by Hadoop MapReduce that provides parallel processing of data present in HDFS. Here each data node works with a part of data which is stored on it which takes less time to process data as compared to when master node alone processes the huge amount of data.

What is Big Data Security?

Big data security is the collective term for all the measures and tools used to safeguard both the data and analytics processes from attacks, theft, or other malicious acts that could harm or negatively affect them. These attacks originate either from offline or online spheres.

The most everlasting problem in any field is security. Security implies that the data or the components which are involved in a network must be safeguarded from any types of attacks and threats. There are risks that are involved when working with an ample amount of data.

Big data security issues [5] [8]:

1. **Distributed frameworks:** Most big data implementations actually distribute huge processing jobs across many systems for faster analysis. Hadoop is a well-known example of open source tech involved in this.
2. **Non-relational data stores:** Think NoSQL databases, which by themselves usually lack security.

3. **Storage:** In big data architecture, the data is usually stored on multiple tiers, depending on business needs for performance vs. cost.
4. **Endpoints:** Security solutions that draw logs from endpoints will need to validate the authenticity of those endpoints, or the analysis isn't going to do much good.

There are several ways an organization can implement security measures to protect its big data[9].
One of the most common security tools is Encryption, a relatively an easy tool that can go a long way. **Encryption** is a process of transformation of plaintext to ciphertext. It is the process of encoding a message or information in such a way that only authorized parties can access it and those who are not authorized cannot. Encrypted data is useless to hackers if they don't have the key to unlock it. Moreover, encrypting the data means that both at input and output, information is completely protected.

Second is building a strong firewall. A ***firewall*** is a network security system which monitors and controls incoming and outgoing network traffic based on predetermined security rules. A ***firewall*** establishes a barrier between a trusted internal network and an untrusted external network. Firewalls are effective at filtering traffic that enters and leaves servers. Organizations can prevent attacks before they happen by establishing strong filters that avoid any third parties or unknown data sources.

Third is use of two-factor authentication (2FA) 2FA adds an additional layer of protection to the authentication process. It requires users to provide any another piece of identifying information in addition to a password. Google brought the advanced form of online security into the mainstream with the launch of multilayered protection for enterprise customers.

Two-factor authentication or two-step verification?
A lot of people think that they are the same thing, but that's not accurate.

There are three types of authentication factors:

1. Password or PIN
2. Mobile phone or a special USB key
3. Fingerprint or other biometric identifier.

As two-factor authentication combines two different factors, two-step verification uses the same factor twice, for example a password and a one-time-code sent via email or SMS. While two-factor authentication is more secured than two-step verification, both are better than relying on a single password.

Fourth is using Virtual Private Networks (VPNs)
A **VPN (virtual private network)**, is a secured tunnel between two or more devices. It uses the Internet as the public backbone to access a secure private network. VPNsprotect private web traffic from snooping, interference, and censorship.VPN technology was developed to allow remote users and branch offices to securely access corporate applications and resources. To ensure security, data would definitely travel through secure tunnels and VPN users would use authentication methods – including passwords, tokens and other unique identification methods – to gain access to the VPN.VPNs work by acting as an intermediary server between you and the site you're visiting. So, if somebody is monitoring your activity, all they will see is the server and not your device. This protects you from unauthorized monitoring and keeps your identity safe.

Fifth is the use of Cryptography which is associated with the process of converting ordinary plain text into unintelligible text and vice-versa. It is a method of storing and transmitting data in a specific form so that only those for whom it is intended can read and process it. Cryptography not only protects information from theft or alteration but can also be used for user authentication. Digital Signatures are based on cryptography. A digital signature is a mathematical technique used to validate the authenticity and integrity of any message, software or digital document.

The digital equivalent of a handwritten signature or stamped seal, a digital signature offers far more intrinsic security, and it is intended to solve the problem of tampering and impersonation in digital communications.

Encryption Algorithm[14]:

This is a Symmetric Key Algorithm. Symmetric encryption has five components [13]:

- **Plain text:** This is the original message that is understandable and is fed into algorithm as input.
- **Encryption algorithm**: It performs various operations and transformations on the plain text.
- **Secret Key:** It is shared between the sender and the receiver and is given as input to the algorithm.
- **Cipher text:** It is the algorithm's output. It is scrambled message and unintelligible that depends on the plain text and the encryption key.
- **Decryption algorithm:** The reverse operation is applied on the cipher text to get back the plain text.

The algorithm works on 64-bit plain text block at a time. Firstly the 64-bit data is divided into 4 portions each of 16 bits (say R1, R2, R3, R4). Thus, R1 to R4 are the inputs for first round of algorithm and there are 8 such rounds. The Key consists of 128 bits. In each round 6 sub keys are generated each of 16 bits [11].

Single Round Details:

There are in total 8 rounds. Steps to be followed in each round are:

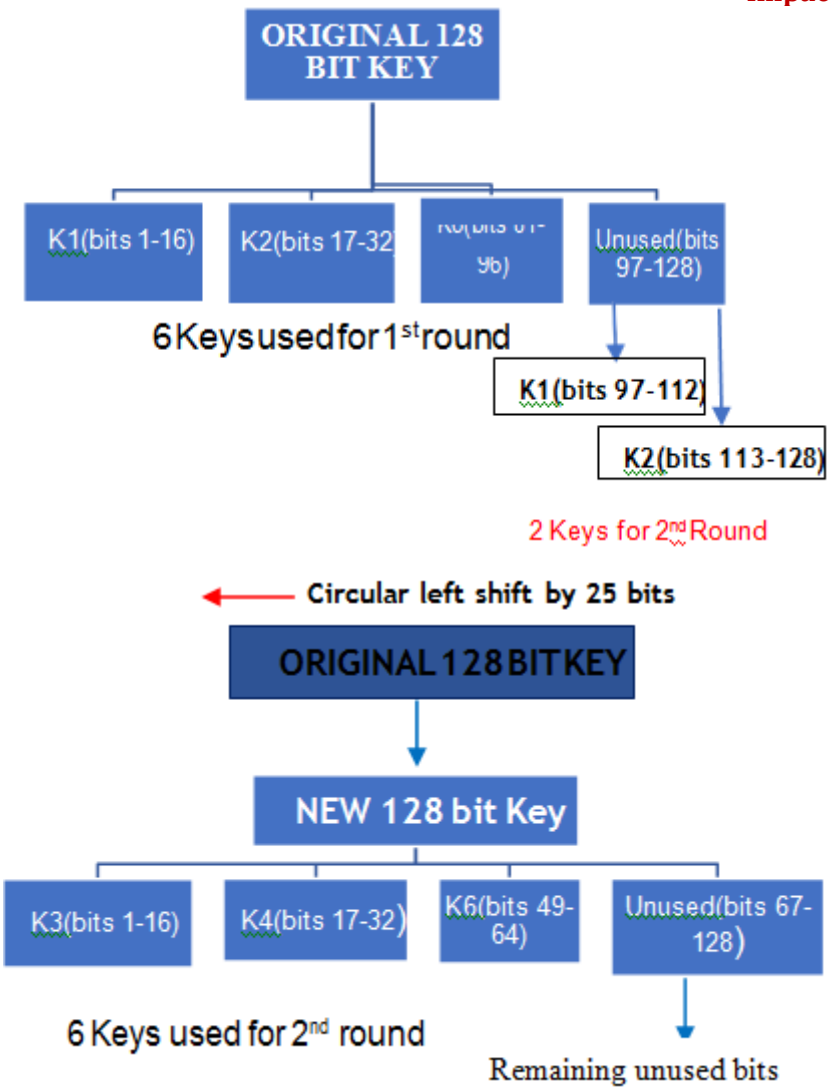
1. R1*K1
2. R2+K2
3. R3+K3
4. R4*K4
5. Step 1 XOR Step 3
6. Step 2 XOR Step 4
7. Step 5 * K5
8. Step 6 + Step 7
9. Step 8 * K6
10. Step 7 + Step 9
11. Step 1 XOR Step 9 R1 →
12. Step 3 XOR Step 9 R2 →
13. Step 2 XOR Step 10 R3 →
14. Step 4 XOR Step 10 R4 →

Now the new inputs to the next round are: R1 R3 R2 R4

Sub-Key generation for a round:

- Since each of the eight round makes use of 6 sub-keys therefore 8*6=48 sub-keys are required for the rounds.
- Also the final output transformation uses 4 sub-keys i.e. 48+4=52 sub-keys in all.
- From an input key 128 bits, these 52 sub- keys are generated.

ISSN 2348 – 8034
Impact Factor- 5.070



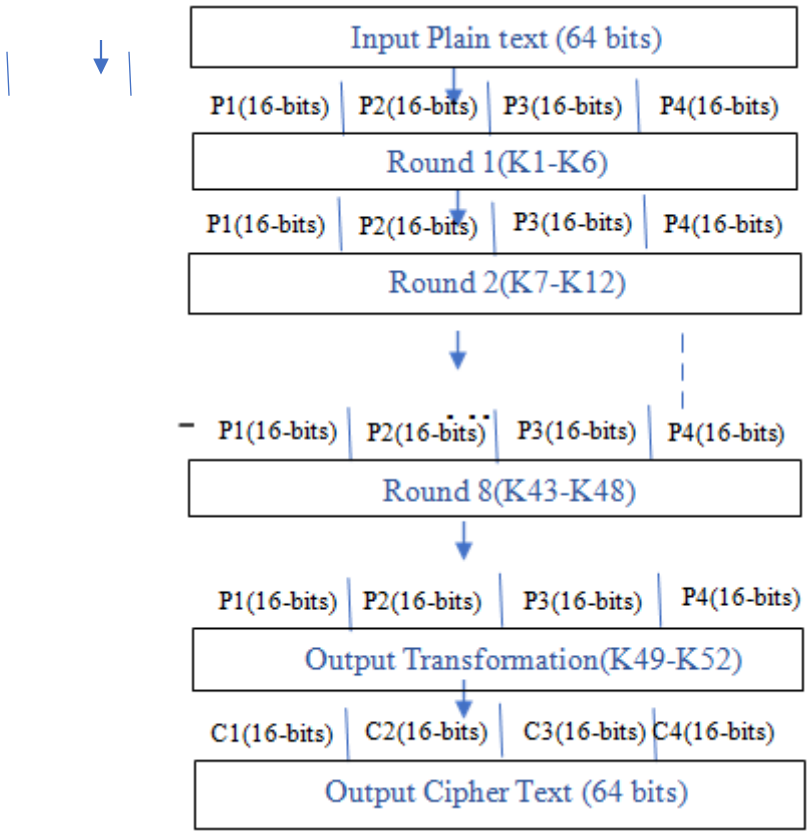
The new key is generated by shifting the bits of the original key circularly to the left afterevery round. As a result, the 26th bit of the original key shifts to the first position and becomes the first bit of the new key and the 25th bit of the original key moves to the last position and becomes 128th bit (after first shift) [12].

Output Transformation:

- It is one-time operation.
- It takes place after the 8th round.
- The key is shifted again left by 25 bits and the [12] first 64 bits of the shifted key are taken for use and used as sub keys K49 to K52 in this phase.

Steps to be followed in this phase:
1. R1*K1
2. R2+K2

3. R3+K3
4. R4*K4



III. CONCLUSION AND FUTURE WORK

The proposed algorithm is a strong block cipher which involves basically only three major operations. Also, the key is made up using 128 bits so it will be less prone to probability of attacks.

In the future this algorithm is going to be enhanced by including more security operations which will strengthen the integrity and confidentiality of data. It can be applied in e-commerce and e-business with slightly different implementation methods.

REFERENCES

1. https://en.wikipedia.org/wiki/Big_data
2. https://www.oracle.com
3. https://imaginenext.ingrammicro.com/.../10- Things-to-Know-About-Big-Data-Securit...
4. hadoop training -eureka playlist 5.www.ijarcsse.com 6.https://en.wikipedia.org/wiki/Apache_Hadoop https://www.sas.com › SAS Insights › Big Data
5. https://www.hack2secure.com/blogs/top-10-big- data-security-challenges
6. https://in.pcmag.com/feature/107583/10-best- practices-for-securing-big-data
7. Challenges and Security Issues in Big Data Analysis. Reena Singh. Kunver Arif Ali. IJIRSET. Volume: 5. Issue: 1. January 2016.
8. www.rroij.com
9. International data encryption algorithm idea-a typical illustration by Sandip Basu, Deptt. Of CS, Asutosh

college, Calcutta university, Kolkata.
10. Introducing an Encryption Algorithm based onIDEA,Osama Almasri1, Hajar Mat Jani2University Tenaga Nasional, College of Graduate Studies, Jalan IKRAM- UNITEN, 43000 Kajang, Selangor, Malaysia University Tenaga Nasional, College of Information Technology, Jalan IKRAM- UNITEN, 43000 Kajang, Malaysia
11. https://en.wikipedia.org/wiki/International_Data_Enc ryption_Algorithm
12. https://goo.gl/Ze1FpX16. S. William,S. Stallings, Cryptography andNetwork Security, Pearson Education, India, 2006.
13. .O. Almasri, H. Mat Jani, Z. Ibrahim, O. Zughoul, "Improving Security Measures of E-Learning Database," International Organization ofScientific Research-Journal of Computer Engineering(IOSR-JCE), 10 (4), pp. 55-62, 2013
14. A. Biryukov, J. Nakahara Jr, B. Preneel, J. Vandewalle, "New Weak-Key Classes of IDEA," In Proceedings of the International Conference on Information and Communications Security (ICICS), pp. 315-326, 2002. [5]
15. Mediacrypt AG, "The IDEA Block Cipher,"cryptonessie.org, 2000. [Online] Available:http://cryptonessie.org [Accessed: Aug. 2, 2013]
16. .H.P. Singh, S. Verma, S. Mishra, "Secure- InternationalData Encryption Algorithm," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2 (2), pp. 780-792, 2013