International Journal for Rapid Research in Engineering Technology & Applied Science
Vol 5 Issue 6 June 2019
ISSN Online:- (2455-4723)

Securing Hadoop: A Survey to investigate need of security in Big Data Processing using Hadoop Ecosystem

Sonal Jain [1]
M. Tech Scholar, CSE Department
BM College of Technology, Indore, India
imsonaljain@gmail.com

Mohit Jain[2]
Assistant Professor, CSE Department
BM College of Technology, Indore, India
bmctmohitcs@gmail.com

ABSTRACT

Big data is used to store bulk amount of data. Hadoop processing system involves large data to deal with and offers scalable and distributed storage. In every minutes and seconds, large data is generated, with the generation of large data they required to be store in a safe and secure manner. As data, leakage is common in Hadoop Distributed File System so security methods need to be implemented in scenario. Existing work uses ARIA and AES algorithm and faces the issue of memory overheads and extra computation time. The drawbacks of existing work are overcome in presented work by replacing AES with Blowfish algorithm and ARIA by RC6. And also serve with high security architecture.

Keywords: **Hadoop, ARIA, Blowfish, Rc6, Security, Big Data**

[1] INTRODUCTION

Big data comprises of data, which is in structured and unstructured format. A massive data with different file formats are stored in Big Data. Hadoop provides with a platform to deal with bulk data by offering them scalable and distributed storage area. Case study on big data says that it is a never-ending deal of data evolution, which is too vast and big. Large data is complex to handle and care of with creating complicated environment. Even security becomes complicated for big data to encrypt due to vulnerable environment for attackers.

Big Data is described in the form of V's:

Volume: Quantity of data, which defines its size.
Variety: Type and nature of data. Data is in which format, either structured or unstructured.
Velocity: Velocity defines the speed of data at which it is generating.
Veracity: Data quality and data value for accurate analysis.

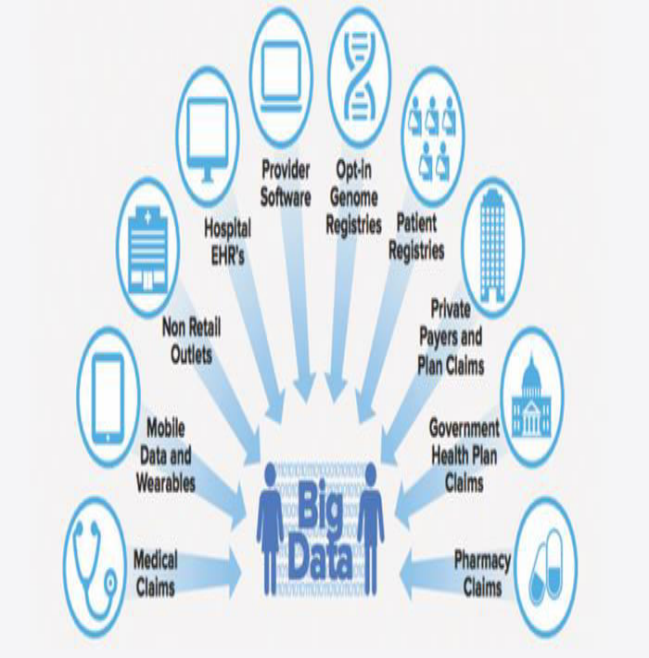

Figure 1: Evolution of Big Data

Hadoop Ecosystem:

Hadoop ecosystem is a framework to solve the issue of big data. Hadoop component together combined to form Hadoop Ecosystem. These components are as:

- MapReduce: is used for distributed processing
- YARN: YARN stands for yet another resource negotiator
- Pig and Hive: Gives SQL data warehouse framework for data query and analysis
- HDFS: Hadoop Distributed File System
- Oozie: Workflow & Job Scheduling
- Zookeeper: Managing cluster & coordinator
- Mahout, Spark: Machine Learning
- HBase: NoSQL database
- Sqoop: Data integration
- Ambari: Management & Monitoring.

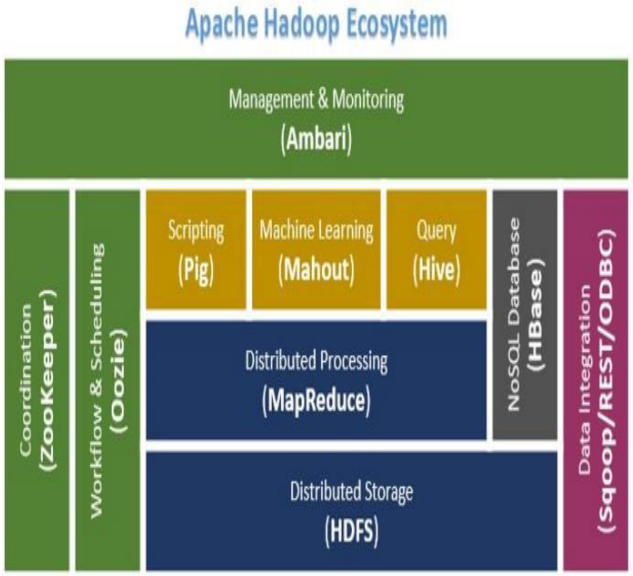

Figure 2. Hadoop Ecosystem

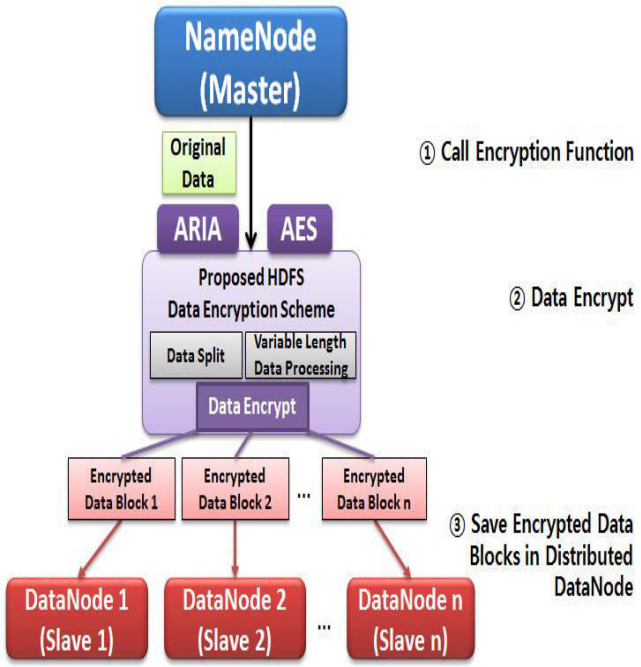

Figure 3: Existing Architecture

ARIA:

ARIA is developed by the researchers of South Korea in South Korea, in the year 2003. They invented this algorithm for security of data and this algorithm is selected as the standard of cryptographic in the year 2004 by the government of Korean in Korea agency for technology and standard.

ARIA algorithm performs XOR operation with block size of 128 bit and key length of 192, 256 bit. It protects and secure data. In presented work, HDFS is used for data storage and scheme designed in such a way that ARIA and BLOWFISH is used to form security architecture.

[2] RELATED WORK

Youngho Song et al. In[1] proposed work using ARIA, which is a Korean government algorithm. He used ARIA with AES algorithm for securing leakage of data. AES is the standard data encryption algorithm, which supports ARIA using HDFS data encryption scheme. Blocks are split and then on it encryption and decryption is performed using ARIA and AES.

S. Ghemawat et al. In [2] described about storage of data in file format. Every format of data is consisted in big data, either structured or unstructured. Big data manages and stores data in a consequent manner. With it author also described Hadoop framework, which comprises of HDFS and MapReduce.

Seon Young Park et al. In[3] abstracted the security of data in Hadoop distributed file system. For this Hadoop requires a proper method to secure data. Security, management, processing and analytics are integrated and comprised in secure method. Some of the factors which security methods consist of is protection of data, visibility, access control, lifecycle, distributed system.

Mathur et al. [4] introduces security algorithm and apply it on plain text for evaluating computation time and different plain text input. Author introduces comparison of different encryption decryption algorithms like AES, RC6, BLOWFISH, DES.

Daesung Kwon et al. In[5] proposed ARIA which is based on SPN_structure, ARIA is a 128 bit block cipher which uses XOR operation with 16 × 16 binary matrix. Using ARIA does not create any duplicacy key issue because it is always applied with XOR.

Zerfos, Petros et al. In[6] elaborated data protection technique like encryption and decryption. HDFS security requirement and data protection can be achieved using some encryption techniques in Hadoop system. This technique for

encryption is best for the security of any data. Author proposed about the system for which security can be established. For the Hadoop-as-service, a secure system is developed, which author proposed it as a Secure Distributed File System for data at rest.

[3] PROBLEM STATEMENT

Security is must whenever talking about data storage. In Hadoop data is stored in HDFS (Hadoop Distributed File System). Process follows as: operating system request hardware for storage and then that dataset is stored in HDFS.

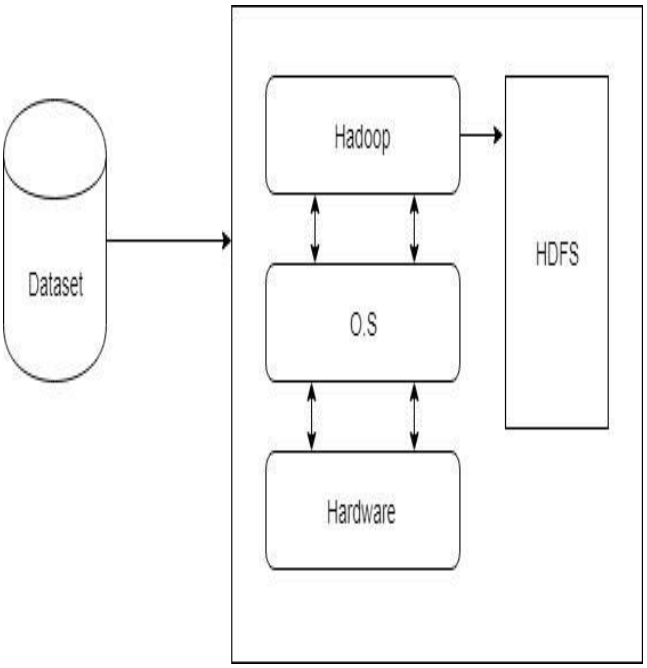

Figure 4: Data storage in Hadoop

Whenever clusters are formed in Hadoop, master node, slave node 1 and slave node 2 is required. They all work as single system, when they are merged the biggest problem evaluate which is the security of data stored in HDFs.

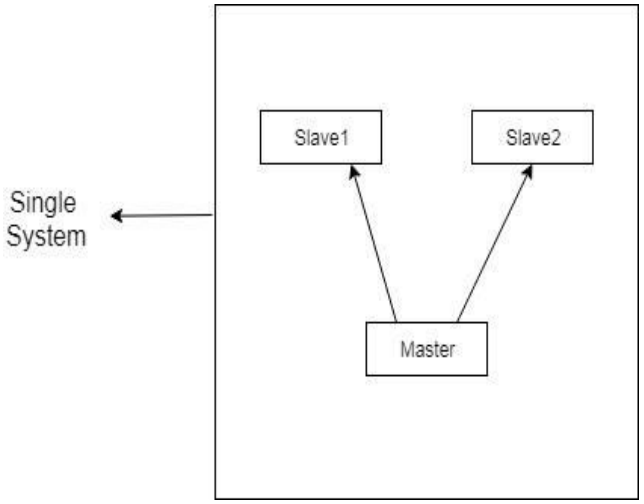

Figure 5: Master-Slave system

When data is stored in HDFS, it acts as a single file system and does not have any security or privacy parameter to secure stored data. Free storage facility is provided by HDFS to store data, as it is general so internal safety is the big question. Internal safety can be explained by example: C1 is the client and C2 is another client. C1 and C2 both fetches data from HDFS, so it might be possible that C2 can grab C1's data. This is based on security of your data by using your own security algorithms.

Boundaries of existing Work:

- Existing system uses AES algorithm, which suffers with memory overhead and extra computation time.
- AES encrypt every block data in the same way.
- HDFS suffers with the issue of security and privacy because data stored in it is not secure.

[4] PROPOSED SOLUTION

Hadoop does not secure your data and if Hadoop do so, then it has its own encryption algorithm to secure data and can decrypt that data using decryption algorithm.

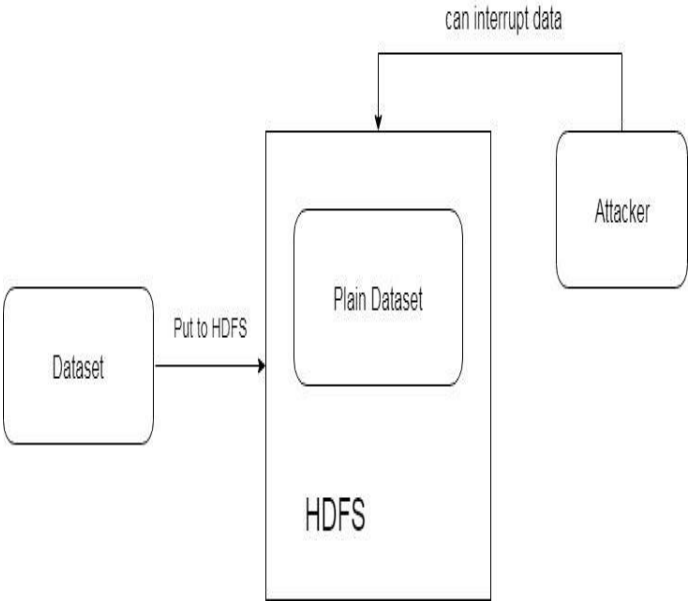

Figure 6: Without Security Architecture

Whenever client uploads data in HDFS, client should definitely use security algorithm, so user fetching that data will get ciphered text.

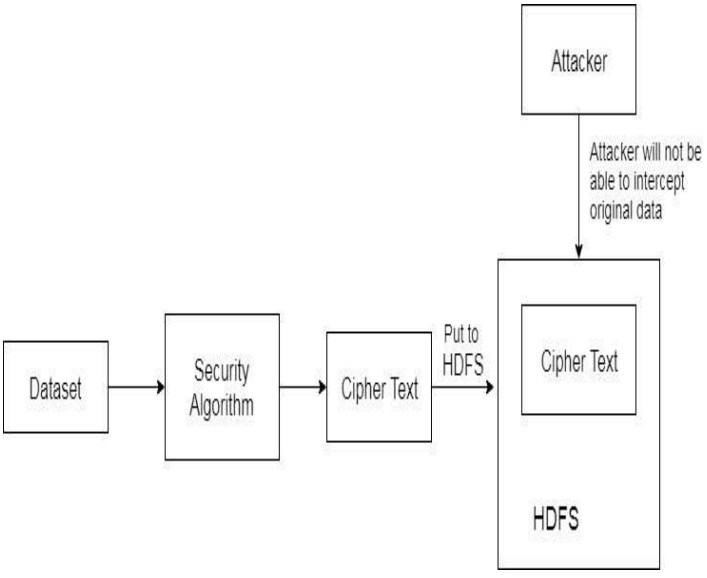

Figure 6: With Security Architecture

Proposed work will replace ARIA and AES with RC6 & Blowfish Algorithms Study of security algorithms address that Blowfish and RC6 can perform better and strong with respect to ARIA and AES.

- BLOWFISH perform better as comparison to AES.
- Key size increase up to 448 bits.
- Improves computation time and memory overheads.

[5] CONCLUSION

Concluded work satisfies the author by replacing AES with Blowfish and ARIA by RC6. As existing work is implemented using AES algorithm with ARIA and face multiple issue of time and memory with security. So Blowfish and RC6 will be implemented in presented work to define high security algorithm using Hybrid Architecture. Security architecture is proposed in solution section where client uploads data in HDFS using security algorithm and in return gets cipher text.

[6] REFERENCES

[1] Youngho Song, Young-Sung Shin, Miyoung Jang & Jae-Woo Chang's "Design and Implementation of HDFS Data Encryption Scheme using ARIA Algorithm on Hadoop" published in IEEE International Conference on Big Data and Smart Computing (BigComp), 2017.

[2] Jeffrey Dean, Sanjay Ghemawat, "MapReduce: Simplified data processing on large clusters," Communications of the ACM, Vol.51, Issue.1, 2008, pp.107-113.

[3] So Hyeon Park and Ik Rae Jeong, "A Study on Security Improvement in Hadoop Distributed File System Based on Kerberos," Journal of the Korea Institute of Information Security and Cryptology, Vol.23, Issue.5, 2013, pp.803-813

[4] Milind Mathur & Ayush Kesarwani's, "Comparison between DES, 3DES, RC2, RC6, & AES" published in NCNHIT 2013.

[5] Daesung Kwon et. al., "New Block Cipher: ARIA" published by National Security Research Institute Korea.

[6] Zerfos, Petros, Hangu Yeo, Brent D. Paulovicks, and Vadim Sheinin. "SDFS: Secure distributed file system for data-at-rest security for Hadoop-as-a-service." In Big Data (Big Data), 2015 IEEE International Conference on, pp. 1262-1271. IEEE, 2015.

[7] Sudheesh Narayanan, "Securing Hadoop: Implement robust end-to-end security for your Hadoop ecosystem," 1st Vol, PACKT Publishing, 2014

[8] Weizhong Zhao, Huifang Ma, Qing He, "Parallel k-means clustering based on mapreduce," In: IEEE International Conference on Cloud Computing. Springer Berlin Heidelberg, Vol.5931 p. 674-679, 2009.

[9] Madhvaraj M Shetty & Manjaiah D.H's "Data Security in Hadoop Distributed File System" published in ICETT 2016.

[10] A. Biryukobv; De Canniere; J.lano; B. Preneel; S. B.Ors's "Security & Performance Analysis of ARIA".