iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** V **Month of publication:** May 2018

DOI: http://doi.org/10.22214/ijraset.2018.5371

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue V, May 2018- Available at www.ijraset.com

Secure Data Storage and Large Data Processing using Hadoop Ecosystem

Sonal jain[1], Asst. Prof. Mohit jain[2]
[1, 2] *B. M. College Of Technology*

Abstract: ***Data is rapidly increasing with great speed now-a-days, and this explosion of data creates issue of security, privacy and handling information leakage. Many of the researchers have given their contribution in solving this issue, but with the change in technology and trend there is a strong need to develop more advanced solution to deal with these kinds of issues of security and privacy. In this work, firstly encryption is performed on large data before storing that data in HDFS, so that, safety of data will be maintained. A security framework with Hadoop ecosystem is proposed. In existing work AES algorithm is proposed, on which observed solution says, that it provides with low performance and computation overheads. This work deals with using Blowfish algorithm by replacing AES to give better solution of the problem.***
Keywords: Blowfish, Data security, Hadoop, Big data, Encryption.

I. INTRODUCTION

A framework called MapReduce is adopted broadly by several organizations for the processing of large volume of data. Thus, issue of processing large datasets is solved using MapReduce. Hadoop environment which works on HDFS, MapReduce, Pig, Hive etc. with creating an ecosystem. It is integrated in Linux for effective computation. A distributed file system is designed called as Hadoop Distributed File System and is also based on java. It is designed with low cost and high fault tolerance with the reliability and scalability of data. It is also beneficial because it provides with replication of data across cluster. No restriction is there for storage of data and provides Support to any format of data which might be structured or any scheme-less.

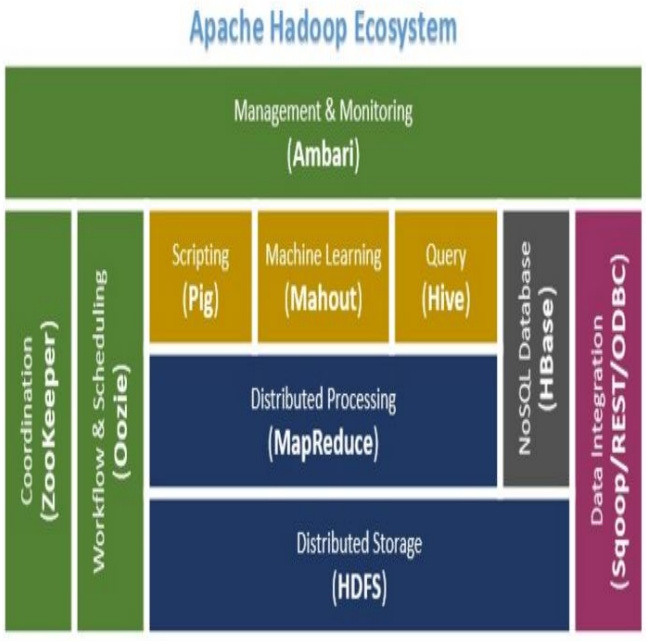


Figure 1.1: Hadoop ecosystem and its components

Big data is the big challenge and is distributed over a distributed environment which make it more complex and vulnerable. Without encryption, any attack can be implemented on data with resulting in security problem. Big data is the integration of huge volume of

data which can be of any form. It has the capability of storage of complex data. With the broad popularity of it organizations and development firms deal in it. Large amount of data is generated in every single second and is categorized into three V's:

A. Volume
B. Variety
C. Velocity.

II. RELATED WORK

A. Study of Base Paper with Diagram

Parmar et. al.[1] introduces about growth of data which is extending rapidly. Thus, processing of this large data and its storage is more complicated and needs different level of advanced algorithm for processing. Valuable data is processed using Hadoop ecosystem. Where security is the significant concern in HDFS because there is no supply of providing privacy and security for the leakage of information. This system uses Kerberos and AES as a security model for data safety.

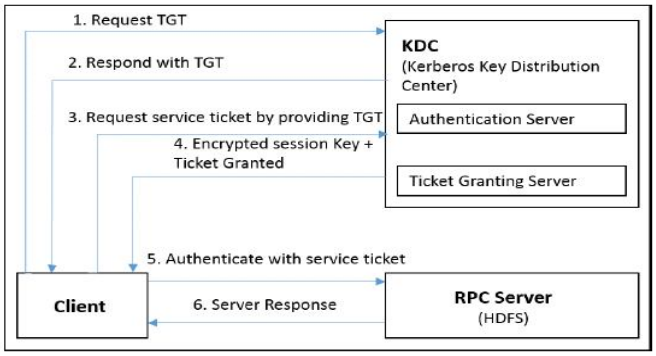


Figure 2.1: Block Representation-1 of Existing Solution

Author observed all the identified security issues and proposes solution to overcome the problem. Vulnerabilities of framework is eliminated by the proposed model

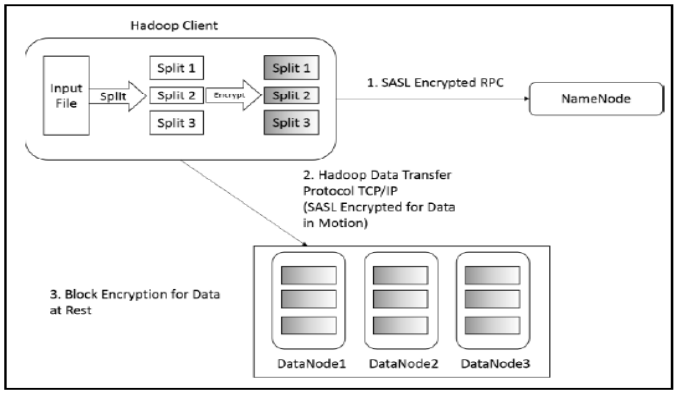


Figure 2.2: Block Representation-2 of Existing Solution

Mathur et al. [2] proposes different security algorithms for data encryption and implemented encryption algorithm which works on computation time of various plain text. Author also compared performance of all the six-significant algorithm.

S. Ghemawat et al. In [3] introduces about big data which can be in any format either structured, unstructured, relational or non-relational. It also consists of complex datasets and these complex dataset produces complexity in dealing with data. Traditional processing of data becomes difficult to handle these types of work.

D. Borthakur et al. In [4] address parallel processing and reliability of distributed data and hardware. Map reduce is also described here which divides data into chunks and works in a parallel way. Also, explained about the working of Hadoop architecture and aims on security issues of Hadoop environment. One way security for the protection of data is encryption. B. Lakhe et al. In [5] described about security of Hadoop which performs parallel processing based on hardware cluster. Chunk data on which parallel task is performed.

III. PROBLEM STATEMENT

Privacy and security of big data is the biggest concern and it should be handled carefully because of the risk of data security and leakage of information. Data is always available in a sharable form and stored in shared storage which is called as HDFS. This is the biggest challenge which is required to be overcome due to the availability of large data in a similar storage. So, for it Big data analysis tool is required which retrieve information and provides with best solutions. Existing work has been analyzed and issue evolve in it generates the problem of leakage of information and insecurity of data. Existing work is based on AES algorithm which when compared proves that blowfish is better in terms of performance. Complete analysis of observed some problem in existing work and is cited below:

A. Primary requirement in Hadoop ecosystem is data encryption because of the protection and safety of data from unauthorized data.
B. If any of the encryption algorithm is used, then it faces the issue of extra overhead and its reductions is important.
C. Performance of Hadoop ecosystem is improved using blowfish in proposed work with replacing AES from existing work.
D. Existing solution needs improvement with the replacement of AES by another algorithm.
E. Key length can also be increase from 128 bit to 448 bit and AES is symmetric key cryptographic with using only single keys.

IV. PROPOSED SOLUTION

A. Proposed Solution

Methodology of proposed work is based on blowfish algorithm with proper system. Where AES is replaced by blowfish for better solution and high performance because AES serves with the issue of low performance. Mining approach will also be implemented in proposed work with using association rule which mines cipher text. Mathur et al. In [2] proposed that blowfish is better in dealing with performance and computation. Also, serves with low memory overhead, as blowfish is symmetric key algorithm which supports different key lengths like 128, 192 and 256 bit. Blowfish encryption will be performed in proposed work, before storing data in HDFS and performing data mining algorithm using mapper class. Mapper class here is used for parallel processing for the mining of ciphered data and performing encryption.

B. System Architecture

System architecture of the proposed work is explained and cited below as:

1) Input data – Large dataset are used which processes big data using Hadoop ecosystem.
2) After input, data is encrypted using blowfish algorithm using required key length with the block size of 64 bits.
3) When data will be encrypted, it will be divided into blocks and after it is stored in HDFS.
4) On block data Apriori algorithm is applied using association rule. Apriori is used for mining of frequent datasets and items in datasets are correlated using association rule mining.
5) At last final recommendation is done depending on Association rule mining.

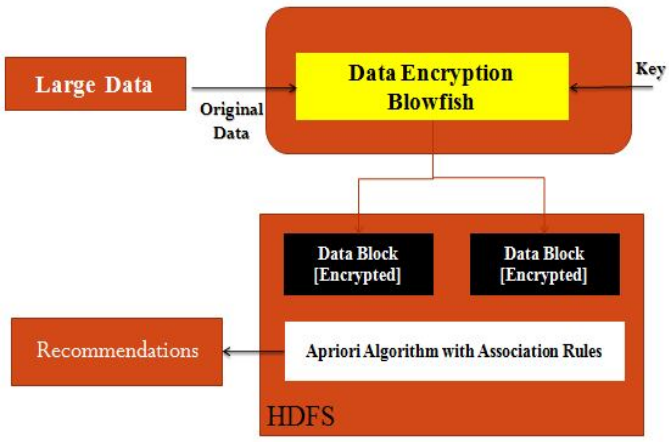


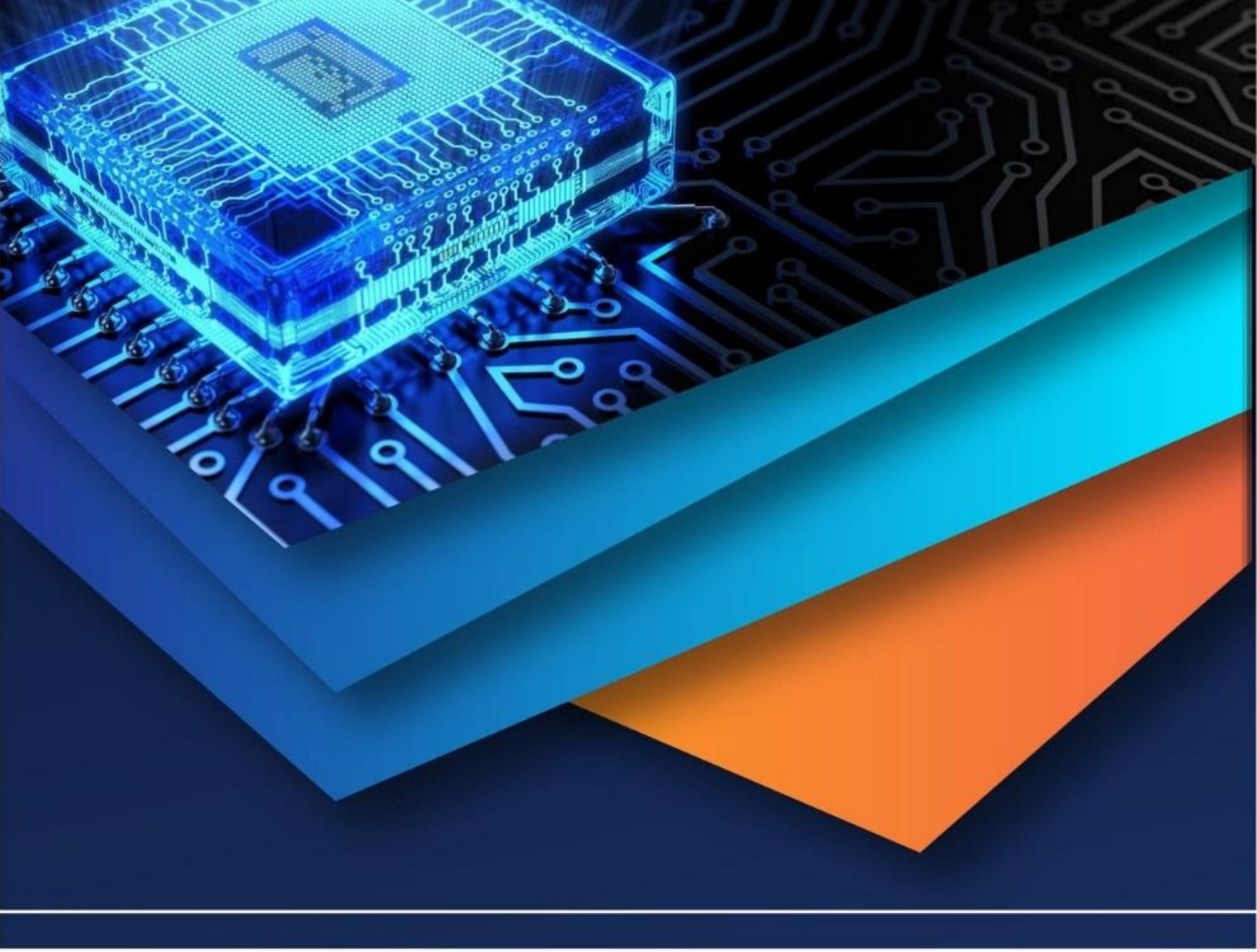
Figure 4.1: System Architecture

V. CONCLUSION

Conclusion of the complete work suggest that there is a strong need to implement security based on confidentiality, access control and authentication. For this encryption of data is required which provides solution over plain text. It serves with permission to authorized person based on authentication. Threshold filter is applied which filter out unwanted efforts. It generates recommendation based on apriori algorithm with association rule mining.

REFERENCES

[1] Raj R. Parmar, Sudipta Roy, Debnath Bhattacharyya, Samir kumar bandyopadhyay, “Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions”. Published on 9 May 2017, pp. 7156-7163, vol. 5, IEEE Access.
[2] [2]. Milind mathur, ayush kesarwani, “comparison between DES, 3DES, RC2, RC6, Blowfish and AES”. Proceedings of National Conference on New Horizons in IT - NCNHIT 2013.
[3] [3]. S. Ghemawat, H. Gobioff, and S.-T. Leung, ‘‘The Google file system,’’ in Proc. 19th ACM Symp. Oper. Syst. Principles (SOSP), 2003, pp. 29–43.
[4] [4]. D. Borthakur, ‘‘The Hadoop distributed file system: Architecture and design,’’ Hadoop Project Website, vol. 11, p. 21, Aug. 2007
[5] [5]. B. Lakhe, Practical Hadoop Security. New York, NY, USA: Apress, 2014, pp.19-46.
[6] [6]. P. P. Sharma and C. P. Navdeti, ‘‘Securing big data Hadoop: A review of security issues, threats and solution,’’ Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 2126–2131, 2014.
[7] [7]. D. J. Bernstein, ‘‘ChaCha, a variant of Salsa20,’’ in Proc. Workshop Rec SASC, 2008, pp. 1–6.
[8] [8]. Seonyoung Park and Youngseok Lee, Secure Hadoop with Encrypted HDFS, Springer-Verlag Berlin Heidelberg in 2013Prof.
[9] [9]. 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.
[10] [10]. Cheng, Zhonghan, Diming Zhang, Hao Huang, and Zhenjiang Qian. "Design and Implementation of Data Encryption in Cloud based on HDFS." International Workshop on Cloud Computing and Information Security (CCIS 2013), pp. 274-277. 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY