HBASE Data Security with AES Algorithm

Chetan Pandey, Amit Juyal, Neeraj Panwar, Aditya Joshi

Abstract: ***Since last decade almost every organization is focusing more on collecting their data (big data) and making analysis of it also applying the concluded valuable outcomes over their organization. The use of smartphones and smart gadgets fasten the gathering of data and enhances the three basic Vs (Volume/ Velocity/ Variety) of big data. This paper focuses on big data security but without fourth V i.e. Value within data, there is no need of securing big data. Perhaps this may be the reason why Hadoop have no security mechanism within its architecture since initially the focus of big data is on the basis of three basic Vs only. With this paper, here authors' try to provide security to big data by using AES algorithm over HBase database. Authors just giving an idea of big data security methodology and for that the main focus of data security is only on valuable contents of the database.***

Keywords: Big Data, AES, Hadoop, HDFS, HBase, NoSQL database

I. INTRODUCTION

Basically Big Data is a term that describes the large amount of data. However every large amount of data is not big data although they are just Data. Big Data is that repository which have continuous storage and different forms of information like social media data which have information in terms of text, icons (emoji), selfies, images, videos and more. As the power of storage is increases (from MB to TB and so on), organizations are now more willing to store all information, either relevant or non-relevant, and analyzing them for insights that lead to better decisions and strategic business moves and promotions. There is one more thing which brings the boom in the database world is the technology. Websites, smartphones, sensors and more all are not only helping human beings but also recording and storing their activities and all those information which we willingly or unwillingly providing to these technologies.

As the power of information gathering is increasing the Volume, which is one of the three Vs of big data (the other two are Velocity and Variety), is becoming more challenging for data management and analytics. Also, according to web study, only around half a percent of all collected data is ever used for analysis. So why to keep all those remaining data? However this part is totally depends on the present and future need of that organization. But the problem remain is that when these data are processed, say simple we fetch a query to just see all stored records or encrypting/decrypting [11][12] the data present in the database, the system need to traverse all that ultra-huge hard drives, ultimately slowing down this whole process.

Revised Manuscript Received on September 25, 2019.

Chetan Pandey , School of Computing, Graphic Era Hill Universtiy, Dehradun, Uttarakhand, India. E-mail: schetanpandey@gmail.com

Amit Juyal, School of Computing, Graphic Era Hill Universtiy, Dehradun, Uttarakhand, India. E-mail: amitjuyal26@gmail.com

Neeraj Panwar, School of Computing, Graphic Era Hill Universtiy, Dehradun, Uttarakhand, India. E-mail: neeraj.pan28@gmail.com

Aditya Joshi, Computer Science Engineering**,** Graphic Era Deemed to be University**,** Dehradun, Uttarakhand, India**.** E-mail: adi.joshi@geu.ac.in

Apart from this, Value, which is one of the important Vs of Big Data, is losing its value in the ocean of big data. This is because most of the data are unstructured and we need strong mining tools and techniques to fetch relevant data out of them and keep them separate within that ocean. Now as the data gaining its Value the risk factor like data breach is also increasing. So again we need some strong security environment which not only providing a secure storage of valuable data but also provide a secure approaches of data analytics.

Keeping the Value, which is one of the important V of big data, we try to first create a big data environment using HBase [15] as the NoSQL database system. HBase is a distributed system which have capability of scalability and most importantly random read and write. HBase is very much helpful for real time work and work effectively with Hadoop's HDFS [16] and Map Reduce features. Focusing on this area authors' tried to make an environment to protect this place of data by using some encryption techniques like AES [11] which is suggested by this paper. For this research work authors' focuses on security of the data which is one of the important shortcomings of big data. However instead of applying security over the entire database the main focus is only on those areas which is important for an organization. Before moving further in this paper, lets first understand what is Value within Big data. As Big data have structured and unstructured both type of data, but out of them there are many fields which are either unrealistic or does not make any impact while data analytic process. Thus Value means that how much the particular data is contributing in the organization growth. Now a days every organizations focusing more on securing big data, since they knew that if somehow the big records goes into wrong hand then there will be big problems form them too. After studying much about different types of security mechanisms, the authors here come to the point that AES is best among all others. Through this paper, it is tried to show a mechanism of securing data on HBase using AES, however the mechanism may be improved in future for better security.

II. RELATED WORK

From the series of research paper it is clear that a data or data set has security issues only when it is important or it have some valuable source of knowledge. Otherwise a data set is secure from all point of views. Big data management and issues related to this is always an interested field of research for researchers and now a day's big data security becomes a new field of curiosity. There are number of research papers focusing big data security and its solution using different parameters like Access control, Threat filtering, Activity monitoring and Alerting.

Some research papers [1] [2] also focusing on the point that the three basic Vs of big data is itself a big issue related to big data security. They also focuses that the architecture of tools and techniques of handling big data are not well secure for example Hadoop uses different layers of handling data from storage at HDFS i.e. last layer to table management at regionserver layer to the top layer Hadoop master layer. Such layers provide chances of hacking of data because there is always a loop hole exists at place where data is migrated from one layer to next layer.

Hadoop, which is platform of storage and processing of big data [3], does not focuses on the security since the beginning of its services, and focuses on this area there are some papers [4] [5] which provides some methodology to achieving security of big data in Hadoop. A paper in year 2016 suggested improved method of applying encryption over the big data, considering Volume for encryption. Paper [5] discuss a required framework for managing and protecting the big data in Hadoop using encryption technique like RSA over HDFS and managing the security services with the help of MapReduce. Another paper [6] published in year 2017 suggested a secure sum protocol which provides a strong security environment. This works is done in two ways suggesting two secure gradient descent algorithms, one is for horizontally partitioned data and other for vertically partitioned data. Further many researcher [7] [8] moved towards NoSQL databases for managing big data. Implementing encryption techniques over them like in a paper [9] author proposes a concept of encryption of a NoSQL database named MongoDB. To encrypt a user's data they utilizes additive homomorphic asymmetric cryptosystem for which author said that they achieve a strong privacy protection. With some experiments, paper [9] compare pre-existed relational database with the authors' concept on the basis of accessing data. It is clear from the later paper that data processing of large amount of data (big data) is handled better in NoSQL databases like MongoDB, HBase and more. Some authors [10] also identifies that existing new technologies like Internet of Things, social networking, usage of cloud storage and more allows one of collect a large amount of data. According to a paper [10] this also brings security and privacy issues related to data and unauthorized person also make use of existing new technologies. Here one more issue is identified by authors that apart from the data, the systems used for collecting data and other data processing process are also vulnerable to cyber attack. The paper [10] covers a brief discussion over some concepts and approaches related to security of data and its privacy. It also covers the risks which are related to big data in terms of volume, variety, velocity, varacity and value. Authors also discussed about different perspectives which are required to have a good solutions regarding the later issues.

This paper further covers Data (Big Data) Security with AES algorithm over HBase Database. Section III will cover scope of authors' proposed methodology. The algorithm of the proposed methodology will be discussed in the Section IV which also cover the details of required tools. Section V will show the outcomes of the algorithm and explain how we can secure the "value" data in a database. The conclusion will be discussed in Section VI which further also include the future scope of the proposed methodology.

III. SCOPE OF RESEARCH WORK

This proposed approach is for providing the security to confidential data by using AES Encryption [11] [12]. Overview of proposed methodology is shown in Figure 1.

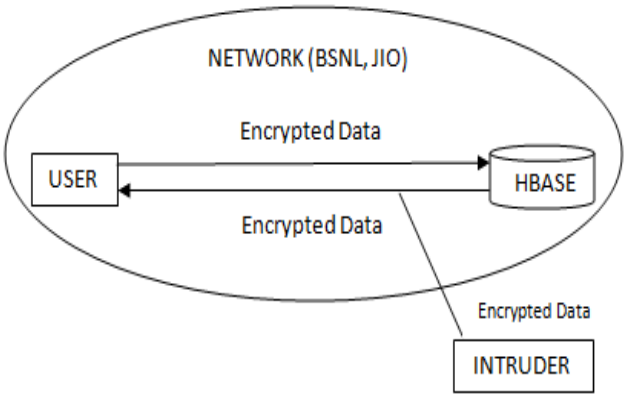


Figure 1: Overview of Proposed Methodology

Within the era of big data in which there exist a lot of sources for collecting data. Many organizations starts making use of NoSQL databases for storing their bulk of data. Here in this paper author making an assumption that if a user uses such a storage or if data is storing in some different locations within a network. Now if the data is in its original form then if somehow intruder gets the data in between of network or from the source of data storage then this become a huge crises for that organization. Here authors are suggesting an approach in which data will be encrypted at user end and then only the data will stored at NoSQL database [9] [15] or move forward in a network. At this time if an intruder somehow gets the data then they will get only the encrypted form of data, which makes original data secure. With the advancements of network and transmission techniques, many organization now making their own datacenters or using cloud storage, transmitting data from one location to another , sharing data and its insight with each other and more. The proposed methodology, which is discussed in next section, is not only securing the valuable data but also making it convenient for transmission.

IV. PROPOSED APPROACH

Before here is more discussion about methodology, authors here focuses on an important point that for this work static data is preferred, real time data will be tested under this security mechanism in future. Authors try to implement encryption and decryption of data over a record of books which having more than 50k rows of data.

The whole methodology is based on the Hadoop HBase and data is secure by using AES algorithm. The programming language used here is java. Below are the recommended versions of tools used for the suggested processing:

Table 1: Recommended Versions of Tools

Java	jdk1.8.0_144
Eclipse	Oxygen Release (4.7.0))
Hadoop	hadoop-2.7.4
HBase	hbase-1.2.6

Apart from these there are two basic hardware requirement i.e. RAM will be 4 GB and HDD will be 500 GB - 1 TB.

Figure 2: Workflow of Proposed Methodology

Before the discussion of algorithm of proposed methodology, there is a need of data which contains a large number of records, here for this research paper authors manage a dataset which contains more than 50k of rows. Firstly it is needed to install Hadoop and over it install HBase [15], since HBase runs over the Hadoop as former uses HDFS [16] technique for data storage and processing. According to its storage structure whole data is stored in HDFS and all operations related to data like read is done by using HDFS. To explain proposed methodology, here authors' opt the AES to encrypt the data but instead of applying this operation over whole record we applied it over a specific column. So our proposed work encrypt one column and store it as encrypted data into the HBase database [15] [16]. If someone see the records in HBase, the encrypted column is show with some encrypted values instead of actual data. However when user enter the encryption key the data is shown in its actual form. Below is the suggested algorithm of suggested methodology.

Table 2: Algorithm of the proposed AES Encryption Security

```
Step 1. Setup Hadoop
Step 2. Setup Hbase
Step 3. Start Hbase
Step 4. Create Hbase table with required fields
Step 5. Initialize text to get 16 bit binary value
Step 6. Define an encryption key
Step 7. Loop till end of file
        Line←Read line
        T←Tokenize line read
        eText←Encrypt required token using AES
        Write the data in Hbase table
       End loop
// Query
Step 8. Input query
Step 9. Loop till end of table
        If query match table column data then
            Display result with encrypted field
            If decrypt then
                Decrypt eText using key
                 Display result
            End If
        End If
        End Loop
Step10. Stop Hbase
Step 11. Stop Hadoop
```

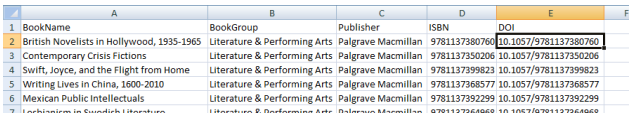
Here we take the following AES steps of encryption for a 128-bit block:

1. Firstly derive the set of round keys from the cipher key.
2. Initialize the state array or the text to be encrypt with the block data (let it plaintext).
3. Add the initial round key to the starting of the state array.
4. Now perform 9 rounds of state manipulation.
5. Perform the 10th round of state manipulation, which is a final round of this process.

Copy the final state array out which is the encrypted data of the given plaintext (let it ciphertext).

V. RESULTS

This proposed security mechanism is working correctly over the large number of records. Authors' tested it under windows environment using Java as programming language. For examining proposed methodology, authors manage a dataset (Figure 3) which contains records of some book having column name as follows: BookName, BookGroup, Publisher, ISBN [17] and DOI. [17] In order to apply and observe the algorithm, here authors assumed DOI as valuable column among others.



	A	B	C	D	E
1	BookName	BookGroup	Publisher	ISBN	DOI
2	British Novelists in Hollywood, 1935-1965	Literature & Performing Arts	Palgrave Macmillan	9781137380760	10.1057/9781137380760
3	Contemporary Crisis Fictions	Literature & Performing Arts	Palgrave Macmillan	9781137350206	10.1057/9781137350206
4	Swift, Joyce, and the Flight from Home	Literature & Performing Arts	Palgrave Macmillan	9781137399823	10.1057/9781137399823
5	Writing Lives in China, 1600-2010	Literature & Performing Arts	Palgrave Macmillan	9781137368577	10.1057/9781137368577
6	Mexican Public Intellectuals	Literature & Performing Arts	Palgrave Macmillan	9781137392299	10.1057/9781137392299

Figure 3: Screenshot of dataset

The algorithm of proposed methodology effectively doing its job and storing it into HBase database. Authors' applied the encryption key into the code for the ease of this research work. In this proposed work, a dataset (Figure 3) is selected and encryption algorithm is applied over the DOI which is considered as a valuable column, also vulnerable to intruders.

For example let a record have DOI = 10.1057/9781137380760. After applying suggested encryption algorithm (Table 2) we get the value as:
Decrypted DOI = ?___??9DX????!K?_M????h?F?-?C?_?

Now this record in HBase will be shown as:

```
BookGroup: Literature & Performing Arts
BookName: British Novelists in Hollywood, 1935-1965
DOI: ?___??9DX????!K?_M????h?F?-?C?_?
ISBN: 9781137380760
Publisher: Palgrave Macmillan
```

So the valuable data is now in its encrypted form and someone if have unprivileged access then only encrypted form of DOI is visible, making valuable information secure. The Figure 4 shows the stored data in the HBase. Here the valuable information i.e. DOI is in its encrypted form.

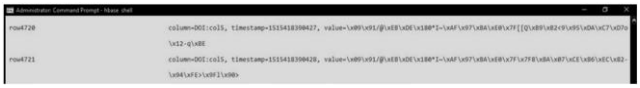


Figure 4: Screenshot showing Data storing in the HBase

Eclipse IDE is used as a coding platform and java is used as a programming language to implement the algorithm of proposed methodology. Figure 5 showing screenshot of it in which data is inserting into HBase along with AES encryption. The values given in the figure is the value which is going to be stored in the HBase. When data moved in HBase the the DOI will be in its encrypted form as shown in Figure 4.

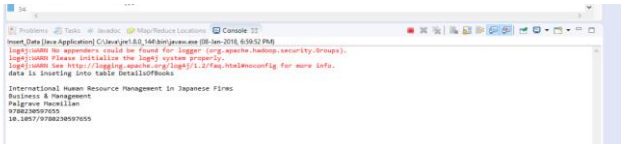


Figure 5: Screenshot showing Data inserting into HBase

Once all data stored in HBase, after this author execute a select query on the basis of ISBN value. This is to make an acknowledgement that whether the data fetch by the select query is in encrypted form or not. Figure 6 shows a screenshot of eclipse, which shows the outcome of the executed query. .

Figure 6: Screenshot showing a Encrypted Data of the user's requested record

VI. CONCLUSION

Although there are so many security techniques to protect the big data. Encryption is one of the most popular and highly accepted tool to not only providing security but also provide privacy by allowing user involvement. DES, Triple DES and AESare some of the encryption technique and among these AES is highly productive. In the above mechanism we encrypted and decrypted the entire database 100% successfully. It is not only easy to understand and implement but also provide much security against unauthorized access. This encryption standard is free to use, has simple hardware requirement, uncomplicated and simple to implement.

However the above used AES encryption is based on key and if somehow the key is hacked then the whole database will be misused or even corrupt. So in order to increase security level of big data it is essential to implement more secure mechanism. A model for securing Big Data is need to be explored that addresses above concerns by implementing some Mathematical Security Model (Encryption and Decryption).

REFERENCES

1. L. Wang, J. Tao, H. Marten, A. Streit, S. U. Khan, J. Kołodziej and D. Chen, "MapReduce Across Distributed Clusters for Data-intensive Applications", IEEE IPDPSW, Shanghai, China, pp 2004 - 2011, 21-25 May 2012\
2. Chao YANG, Weiwei LIN, Mingqi LIU, "A Novel Triple Encryption Scheme for Hadoop-based Cloud Data Security", IEEE EIDWT, 9-11 Sept. 2013, Xi'an, China, pp 437 - 442
3. Atif Mohammad, Hamid Mcheick, Emanuel Grant, "Big Data Architecture Evolution: 2014 and Beyond", ACM MSWiM, Sept 21-26, 2014, Montreal, QC, Canada, pp 139-144 Atif Mohammad, Hamid Mcheick, Emanuel Grant, "Big Data Architecture Evolution: 2014 and Beyond", ACM MSWiM, Sept 21-26, 2014, Montreal, QC, Canada, pp 139-144
4. Karim Abouelmehdi, Abderrahim Beni-Hssane, Hayat Khaloufi, Mostafa Saadi, "Big Data Emerging Issues: Hadoop Security and Privacy", IEEE ICMCS, 29 Sept.-1 Oct. 2016, Marrakech, Morocco, pp 731 - 736
5. Arunima Dubey, Satyajee Srivastava, "A Major Threat To Big Data - Data Security", IEEE ICCCA, 29-30 April 2016, Noida, India, pp 60 - 64
6. Ibtissam Ennajjar, Youness Tabii, Abdelhamid Benkaddour, "Securing Data in Cloud Computing by Classification", ACM BDCA, March 29-30, 2017, Tetouan, Morocco, pp 493-498
7. Shagu.a Mehnaz, Gowtham Bellala, Elisa Bertino, "A Secure Sum Protocol and Its Application to Privacy-preserving Multi-party Analytics", ACM SACMAT, June 7 2017, NY, USA, pp 219-230
8. John Carlo Bertot, Heeyoon Choi, "Big Data and e-Government: Issues, Policies, and Recommendations", ACM DGR, June 17, 2017, NY USA, pp 1-10
9. Guowen Xu, Yan Ren, Hongwei Li, Dongxiao Liu, Yuanshun Dai, Kan Yang, " CryptMDB: A Practical Encrypted MongoDB over Big Data", IEEE ICC, 21-25 May, 2017, Paris, France, pp 1-6
10. Elisa Bertino, Elena Ferrari (2018). Big Data Security and Privacy (vol. 31), Springer.
11. About Advanced Encryption Standard [Online]. Available: https://en.wikipedia.org/wiki/ Advanced_Encryption_Standard
12. About and Working of AES [Online]. Available: http://searchsecurity.techtarget.com/ definition/Advanced-Encryption-Standard
13. Cryptograhy Tutorial [Online]. Available: https://www.tutorialspoint.com/cryptography/ advanced_encryption_standard.htm
14. Encryption Process [Online]. Available: http://etutorials.org/Networking/ 802.11 + security.+ wi-fi + protected + access + and + 802.11i / Appendixes/Appendix + A .+ Overview+of+the+ AES + Block + Cipher / Steps +in+the+AES+Encryption+Process/
15. HBase Overview [Online]. Available: http://moi.vonos.net/bigdata/hbase/
16. HDFS in HBase [Online]. Available: http://www.aosabook.org/en/hdfs.html
17. DOI and ISBN [Online]. Available: https://www.doi.org/factsheets/ISBN-A.html