International Journal of Engineering & Technology, 7 (2.15) (2018) 54-57

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Design and Implementation of Data-at-Rest Encryption for Hadoop

Siti Hanisah Kamaruzaman*, Wan Nor Shuhadah Wan Nik, Mohamad Afendee Mohamed, Zarina Mohamad

Faculty of Informatics and Computing, University Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia
**Email: 038086@putra.unisza.edu.my*

Abstract

The manuscript should contain an abstract. The security aspects in Cloud computing is paramount in order to ensure high quality of Service Level Agreement (SLA) to the cloud computing customers. This issue is more apparent when very large amount of data is involved in this emerging computing environment. Hadoop is an open source software framework that supports large data sets storage and processing in a distributed computing environment and well-known implementation of Map Reduce. Map Reduce is one common programming model to process and handle a large amount of data, specifically in big data analysis. Further, Hadoop Distributed File System (HDFS) is a distributed, scalable and portable file system that is written in java for Hadoop framework. However, the main problem is that the data at rest is not secure where intruders can steal or converts the data stored in this computing environment. Therefore, the AES encryption algorithm has been implemented in HDFS to ensure the security of data stored in HDFS. It is shown that the implementation of AES encryption algorithm is capable to secure data stored in HDFS to some extent.

Keywords: *Encryption; Hadoop; Map-Reduce; Cloud Computing.*

1. Introduction

The security of services in cloud computing is one of the most challenging topics and it is a cloud computing's core technology that is currently becomes the focus of computing and information technology era. Moreover, in every second the amount of data is drastically increases day by day. It is more apparent in faster development of the internet, Internet of Things and Cloud Computing where it led to the rapid growth of data in almost every industry and business area [6]. The development of Big Data had attracted attention from various fields around the world. Big data can be found in three forms, i.e. structured, unstructured and semi-structured [6].

Hadoop is used for big data analysis. It is an open source software framework that allows the distributed processing of big data sets across clusters of computers using simple programming language [6]. It supports large data sets storage and processing in a distributed computing environment. Two main components of Hadoop are 1) Hadoop Distributed File System (HDFS) and 2) Map Reduce. HDFS is a repository of large data sets in Hadoop, while the Map Reduce is a distributed programming model that is responsible to process and handle a vast volume of data stored in Hadoop.

While the data that has been analyzed by Hadoop is stored in HDFS, the security of these data may not be promised. These data may be stolen, modified or converted by intruders. In that case, the encryption needs to be implemented to the data in HDFS to secure these data to some extent.

Therefore, this paper focuses on two main objectives, i.e. 1) to study the architecture of Hadoop and the current implementation of encryption technique for static data (“data-at-rest”) stored in HDFS using AES algorithm, and 2) to test and evaluate the successfulness of AES algorithm in HDFS in simpler forms. By doing these, the prepparation of implementing more complicated or higher level of security measures (specifically in implementing other variants of encryption algorithm) can be done in future.

2. Related Works

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment [3]. The architecture allows running applications on systems which may involve thousands of nodes with thousands of terabytes of data [3]. To be specific, Hadoop consist of several components that are communicated to each other to make up the ecosystem. The main component is the Hadoop kernel, which consists of Map Reduce and the Hadoop distributed file system (HDFS). This kernel is supported by other related components such as Apache Hive, HBase, Oozie, Pig and Zookeeper [3]. Undoubtedly, encryption ensures confidentiality and privacy of user information and it secures the sensitive data in Hadoop [3]. Hadoop did not include basic controls for data protection and most third-party tools could not scale along with NoSQL and so were little use to developers [4].

According to the researcher in [3], static data can be secured in two ways. First, a complete data file that will be stored in Hadoop is encrypted beforehand; second, data is encrypted once they are loaded into Hadoop system. Currently, the implementation of Hadoop, specifically the HDFS, supports the AES encryption algorithm, i.e. the OS level encryption for data at rest. Meanwhile, Zookeeper, Oozie, Hive, HBase, and Pig do not offer data at rest encryption solution but for this component encryption can be implemented via custom encryption techniques or third party tools [3].

Further in [1], the researcher compared the Apache Spark and Apache Hadoop. They concluded that Spark focuses on simplifying the complexity of processing compute-intensive tasks with high volumes of real-time or archived data, both structured and unstructured. It integrates relevant complex capabilities such as machine learning and graph algorithm. That is, by using Spark, Big Data processing can be run over hundreds, thousand, or even tens of thousands of machines in a cluster is merely a configuration change. Comparatively, Apache Spark is not replacing to Hadoop but it is one of the alternatives to Hadoop.

In relation to Cloud computing, where data security is a compulsory, research in [2] provides an improved version of Hadoop, which is capable to establish strong mutual authentication by using Kerberos. In this research, efforts have been done in order to ensure data security in Hadoop-based cloud data storage. A triple combination of HDFS files encryption has been proposed and implemented. DEA (Data Encryption Algorithm) is first used to encrypt the file before the data key encryption is done by using the RSA algorithm. Later, the user's RSA private key is encrypted using IDEA (International Data Encryption Algorithm) [2].

3. Data-at-Rest encryption for Hadoop

In order to achieve the objectives of this paper, the development and testing processes takes advantage of the evolutionary/iterative method. This model is less costly when the alteration of project requirements is needed. By using this method, the testing and debugging processes are easier to be handled when small iterations are involves.

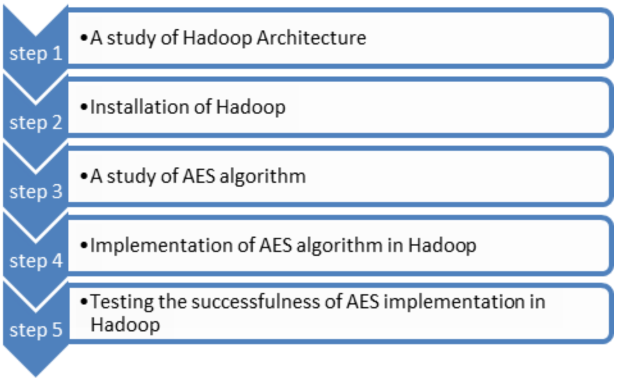

step 1
•A study of Hadoop Architecture
step 2
•Installation of Hadoop
step 3
•A study of AES algorithm
step 4
•Implementation of AES algorithm in Hadoop
step 5
•Testing the successfulness of AES implementation in Hadoop

Fig. 1: Project Flow

Fig. 1 shows the five main steps involved in order to achieve the objective of this paper. Hadoop architecture is first study in the first step. Then, the installation of Hadoop in the server has been done in the second step. Later, when Hadoop is successfully running in the server, the AES encryption algorithm is studied before its implementation and its successfulness can be tested in Hadoop.

Hadoop is installed in a virtual machine (VM) which acts as a Hadoop server. Oracle VM VirtualBox is used in creating the virtual machine where Hadoop architecture is studied. By doing this, the components or framework of Hadoop can be studied in a less cost. Hadoop architecture consist of the Hadoop kernel, Map Reduce and Hadoop Distributed File System (HDFS) and a number of related component such as Apache Hive, HBase, Oozie, Pig and Zookeeper [3].

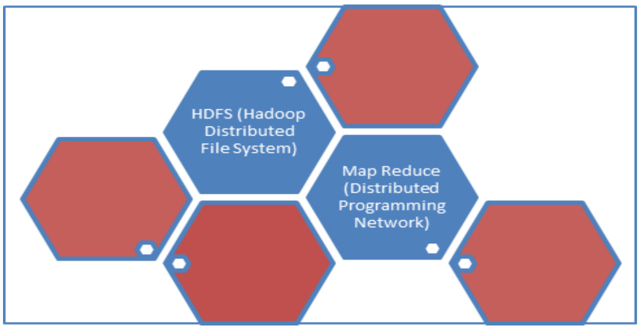

HDFS (Hadoop Distributed File System)
Map Reduce (Distributed Programming Network)

Fig. 2: Main Components of Hadoop

HDFS is a highly faults tolerant distributed file system that is responsible for storing data on the cluster while Map Reduce is a powerful parallel programming technique for distributed processing of vast amount of data on clusters [3]. Besides that, HBase is a column oriented distributed NoSQL database for random read/write access [3]. Meanwhile, Pig is a high level data programming language for analysing data of Hadoop computation [3]. In addition, Hive is a data warehousing application that provides a SQL like access and relational model while Sqoop is a project for transferring or importing data between relational databases and Hadoop [3]. Besides that, Oozie is an orchestration and workflow management for dependent Hadoop jobs [3].

Data at rest in HDFS is encrypted by using AES encryption algorithm. The study of AES algorithm for encryption is needed before it can be implemented in HDFS. If the intruders have the correct key to decrypt the data, the encrypted data will be decrypt into plaintext.

Lastly, the successfulness of the implementation of AES encryption has been tested. The testing stage of the framework must be performed in order to detect any defect that can only be found when it is test in the operational environment.

In this project, AES (Advanced Encryption Standard) encryption technique has been implemented. Encryption is the process of encoding data in such a way that only authorized users can decode and use the data which is self-defensive and enhances data security [1]. Decryption is just the reverse of encryption. The receiver transforms sender's cipher text into a meaningful text known as plaintext [1].

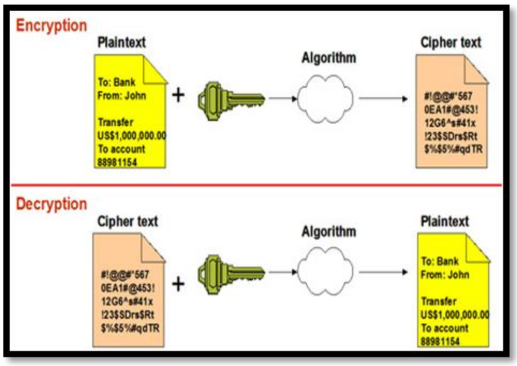


Fig. 3: Encryption and Decryption Techniques

Symmetric encryption is used to encrypt more than a small amount of data. During both the encryption and decryption, the process of symmetric key is used. The key to encrypt the data must be used to decrypt a particular piece of cipher text [7].

AES (Advances Encryption Standard) is the symmetric cryptosystem that uses the same key to make an encryption and decryption. AES algorithm is not only for security but also for great speed [5]. The goal of every encryption algorithm is to make it as difficult as possible to decrypt the generated cipher text without using a key [7].

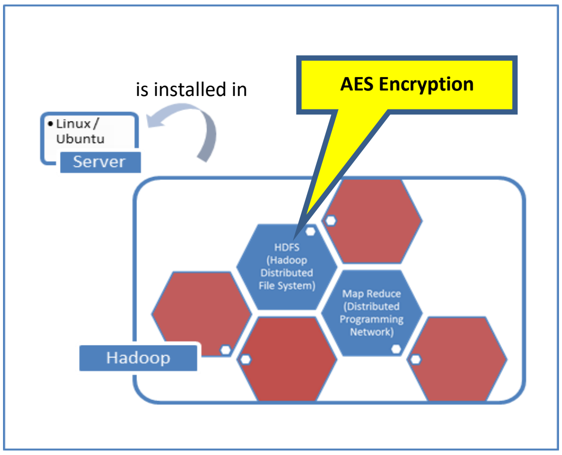
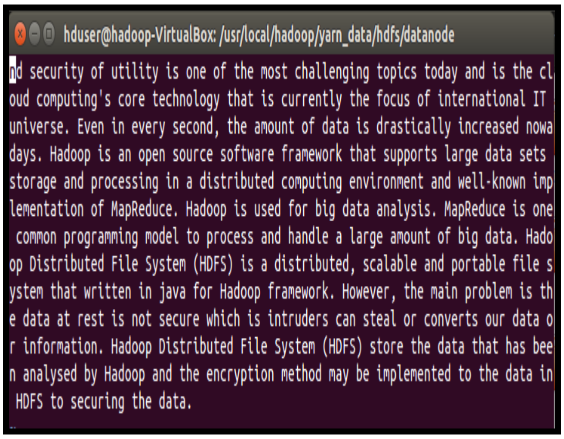

is installed in
AES Encryption
• Linux / Ubuntu
Server
HDFS (Hadoop Distributed File System)
Map Reduce (Distributed Programming Network)
Hadoop

Fig. 4: Project Framework

Based on the framework of the project shown in Fig. 4 above, the server is accessed and controlled by admin. Next, open source software, i.e. Hadoop Apache is installed in the server and some configurations in Hadoop are needed in order to ensure Hadoop is running correctly. Besides that, the encryption method has been implemented in the component of Hadoop which is in HDFS (Hadoop Distributed File System) in the datanode. The data at rest in HDFS will be encrypted by AES encryption algorithm at the datanode. If the intruders have the correct key to decrypt the data, the encrypted data will be decrypt into plaintext.

4. Results and discussions

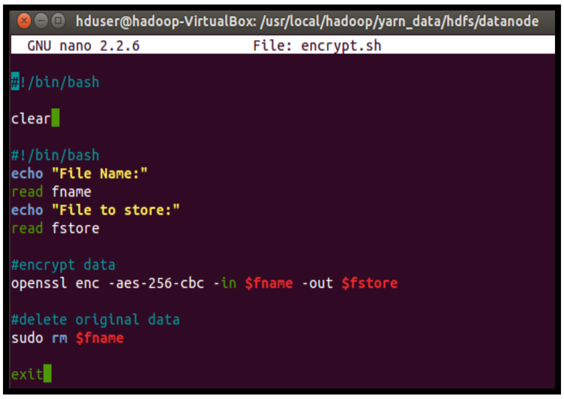
For the testing and result, the file called "data" is created and saved in the datanode file in HDFS.



```
hduser@hadoop-VirtualBox: /usr/local/hadoop/yarn_data/hdfs/datanode
Id security of utility is one of the most challenging topics today and is the cl
oud computing's core technology that is currently the focus of international IT
universe. Even in every second, the amount of data is drastically increased nowa
days. Hadoop is an open source software framework that supports large data sets
storage and processing in a distributed computing environment and well-known imp
lementation of MapReduce. Hadoop is used for big data analysis. MapReduce is one
 common programming model to process and handle a large amount of big data. Hado
op Distributed File System (HDFS) is a distributed, scalable and portable file s
ystem that written in java for Hadoop framework. However, the main problem is th
e data at rest is not secure which is intruders can steal or converts our data o
r information. Hadoop Distributed File System (HDFS) store the data that has bee
n analysed by Hadoop and the encryption method may be implemented to the data in
 HDFS to securing the data.
```

Fig. 5: Original Data

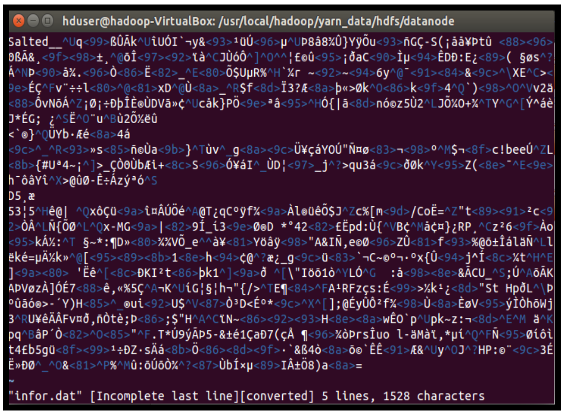
Fig. 5 shows the content of the original data that was created in the HDFS.



```
hduser@hadoop-VirtualBox: /usr/local/hadoop/yarn_data/hdfs/datanode
  GNU nano 2.2.6              File: encrypt.sh

#!/bin/bash

clear

#!/bin/bash
echo "File Name:"
read fname
echo "File to store:"
read fstore

#encrypt data
openssl enc -aes-256-cbc -in $fname -out $fstore

#delete original data
sudo rm $fname

exit
```

Fig. 6: Script for encryption process

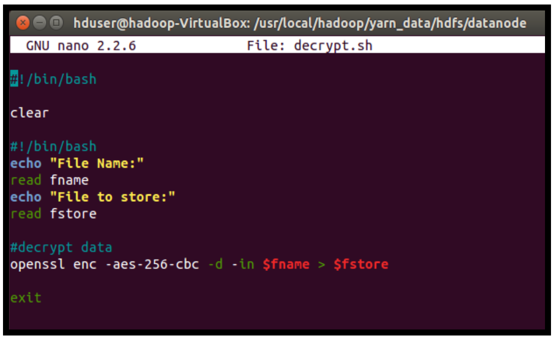
After that, the scripting for encryption process is created and is named as encrypt.sh. In these scripting, the command will read the file name that needs to be encrypted together with the location to store this file when it has been encrypted. Advanced Encryption System (AES) encryption has been used to encrypt the content of the file. AES encryption is a symmetric cryptography that use the same key to make an encryption and decryption. AES encryption not only for security but also for great speed and 256-bit key is use because 256-bit key is stronger than 128-bit key and more difficult to break. The larger the size of key, the more secure the encryption. In this scripting, we also write the command to delete the original data.



```
hduser@hadoop-VirtualBox: /usr/local/hadoop/yarn_data/hdfs/datanode
Salted__^Uq<99>ßÛĀk^UîUÓI`¬y&<93>¹üÚ<96>µ^UÞ8â8¾Û}YÿÕu<93>ñGÇ-S(¡ȁå¥Þtû <88><96>
0ßĀ&¸<9f><98>±¸^@ōÎ<97><92>ĩà^CJÙóÔ^]^O^^¦£©û<95>¡ðaC<90>Ìµ<94>ÊDĐ:E¿<89>( §øs^?
Á^NÞ<90>ǻ¾.<96>Ò<86>Ë<82>_^E<80>Õ$UµR%^H`¼r ~<92>~<94>6y^@¯<91><84>&<9c>^\XE^C><
9e>ÉÇ^Fv¨÷÷l<80>^@<81>xD^@Ù<8a>_^R$f<8d>Ï3?Æ<8a>þ«>Øk^O<86>k<9f>4^Q`)<98>^O^Vv2ä
<88>ÔvNōÁ^Z¡Ø¡÷ÐþÎÈ®ÙDVā»¢^Ucák}PÖ<9e>ªâ<95>^HÓ{|ā<8d>nó®z5Ù2^LJÕ¼O+¾^TY^G^[Ý^áè
J*ÉG; ¿^SË^O¨u^Bù2Õ½ëû
<`®}^QÜYb·Æé<8a>4á
<9c>^_^R<93>»s<85>ñ©Ùa<9b>}^Tùv^_g<8a><9c>Ü¥çáYOÚ"Ñ¤ø<83>¬<98>º^M$¬<8f>c!beeÚ^ZL
<8b>{#Uª4~¡^]>_ÇÒ0ÙbÆi+<8c>S<96>Ó¥áI^_ÙD¦<97>_j^?>qu3á<9c>ðØk^Y<95>Z(<8e>¯^E<9e>
h¯ôâYî^X>@ûØ-Ê÷Ázýªó^S
D5¸æ
53¦5^Hê@| ^QxôÇü<9a>î¤ÂÚÖé^A@T¿qCºÿf¾<9a>Àl®üêÕ$J^Zc%[m<9d>/CoË=^Z"t<89><91>²c<9
2>ÒÂ^LÑ{ÕØ^L^Qx-MG<9a>|<82>9Í_í3<9e>Ø®D *°42<82>£Ëpd:Ù{^VB¢^Mǻ¢¤};RP¸^Cz²6<9f>Ào
<95>kÁ½:^T §~*:¶D»<80>¾¾VÕ_e^^à¥<81>Yöâÿ<98>"A&IÑ,e®Ø<96>ZÛ<81>f<93>%@ō±ÎálāÑ^Ll
ëké=µÄ½k»^@[<95><89><8b>1<8e>h<94>¢@^?æ¿_g<9c>ü<83>`¬C~®º¬·ºx{Û<94>j^Í<8c>¼t^H^E
]<9a><80> 'Ëê^[<8c>ÐKI²t<86>þk1^]<9a>ð ^[\"Iöō1ò^YLÓ^G  :ǻ<98><8e>&ĀCU_^S;Ú^AōĀK
AÞVøzÀ]ÓÉ7<88>ê,«%5Ç^A¬K^UíG¦§¦h¬"{/>^TE¶<84>^FA¹RFzçs:É<99>>¼k¹¿<8d>"St HpðL^\Þ
ºûāó®>-´Y)H<85>^_®ui<92>U$^V<87>Ò³D<Éº*<9c>^X^[];@ÉyÛÔ²f¾<98>Ù<8a>ÈøV<95>ýÌÒhōWj
3^RU¥êÄÂFv¤ð,ñÒtè;Þ<86>;$"H^A^CĩN~<86><92><93>H<8e><8a>wÊO`p^Upk~z:¬<8d>^E^M ä^K
pq^BâP´Ò<82>^O<85>"^F.T*Ú9ýĀÞ5-&±é1ÇaÐ7(çÅ ¶<96>%òÞrsÎuo l-äMàĩ,*µí^Q^FÑ<95>Øíôì
t4£b5gü<8f><99>¹÷ÐZ·sÄá<8b>Ö<86><8d><9f>·`&ß4ò<8a>ō®`ÊÊ<91>Æ&^Uy^OJ^?HP:®¨<9c>3É
Ë»ÐØ^_^O&<81>^P%^Mû:ōÚōÒ¾^?<87>ÙbÍ×µ<89>IÂ±Ö8)a<8a>=
~
"infor.dat" [Incomplete last line][converted] 5 lines, 1528 characters
```

Fig. 7: Encrypted text

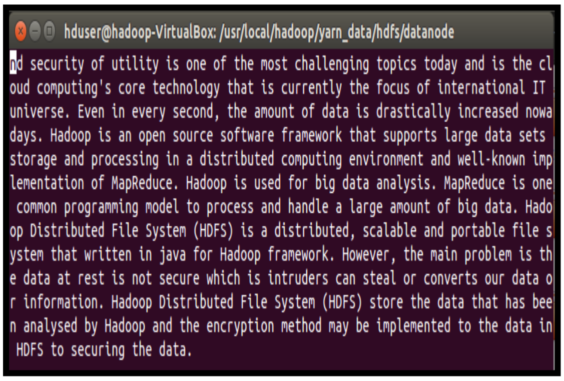
Fig. 7 shows the result after the encryption process has been done. The content in the file become as cipher text. That is, the content of the file cannot be understood until this file is decrypted by using the same key.



```
hduser@hadoop-VirtualBox: /usr/local/hadoop/yarn_data/hdfs/datanode
  GNU nano 2.2.6              File: decrypt.sh

#!/bin/bash

clear

#!/bin/bash
echo "File Name:"
read fname
echo "File to store:"
read fstore

#decrypt data
openssl enc -aes-256-cbc -d -in $fname > $fstore

exit
```

Fig. 8: Script for decryption process

After that, the scripting of decryption is created and is named as decrypt.sh. In these scripting, the command will read the file name that needs to be decrypted together with the location to store the file that was decrypted.



```
hduser@hadoop-VirtualBox: /usr/local/hadoop/yarn_data/hdfs/datanode
Id security of utility is one of the most challenging topics today and is the cl
oud computing's core technology that is currently the focus of international IT
universe. Even in every second, the amount of data is drastically increased nowa
days. Hadoop is an open source software framework that supports large data sets
storage and processing in a distributed computing environment and well-known imp
lementation of MapReduce. Hadoop is used for big data analysis. MapReduce is one
 common programming model to process and handle a large amount of big data. Hado
op Distributed File System (HDFS) is a distributed, scalable and portable file s
ystem that written in java for Hadoop framework. However, the main problem is th
e data at rest is not secure which is intruders can steal or converts our data o
r information. Hadoop Distributed File System (HDFS) store the data that has bee
n analysed by Hadoop and the encryption method may be implemented to the data in
 HDFS to securing the data.
```

Fig. 9: Decrypted text

Fig. 9 shows the result after the decryption process is done in order to retrieve the original data.

5. Conclusions

In conclusion, we can conclude that data security is one of the main issues in Big Data. In this project, Advanced Encryption Standard (AES) encryption has been implemented to prevent unauthorized user or intruders from altering the static data stored in Hadoop system. AES will encrypt the content of the file so that only authorized user can open that file. This project has been tested and the implementation of AES encryption at datanode in HDFS was successfully generated. Therefore, this project has proved that the encryption method implemented in Hadoop is properly functioning to encrypt the content of the file.

References

[1] https://www.researchgate.net/publication/301887194_Efficient_Hybrid_MAES_Encryption_Algorithm_for_Mobile_Device_Data_Security_at_Rest_in_Cloud_Environment.

[2] Yang, C., Lin, W., & Liu, M. (2013, September). A novel triple encryption scheme for Hadoop-based cloud data security. In *Emerging Intelligent Data and Web Technologies (EIDWT), 2013 Fourth International Conference on* (pp. 437-442). IEEE.

[3] Sharma, P. P., & Navdeti, C. P. (2014). Securing big data Hadoop: a review of security issues, threats, and solution. *Int. J. Comput. Sci. Inf. Technol*, *5*.

[4] https://securosis.com/assets/library/reports/Securing_Hadoop_Final_V2.pdf

[5] Padmavathi, B., & Kumari, S. R. (2013). A survey on performance analysis of DES, AES and RSA algorithm along with LSB substitution. *Int. J. Sci. Res*, *2*(4), 170-174.

[6] Pol, U. R. (2016). Big Data Analysis :Comparision of Hadoop MapReduce and Apache Spark Big Data Analysis : Comparision of Hadoop MapReduce and Apache. *International Journal of Engineering Science and Computing*, *6*(6), 6389–6391. https://doi.org/10.4010/2016.1535

[7] Microsoft [Online] Available: https://msdn.microsoft.com/enus/library/windows/desktop/aa381939(v=vs.85).asp

[8] Cohen, J., & Acharya, S. (2013, December). Towards a trusted hadoop storage platform: Design considerations of an aes based encryption scheme with tpm rooted key protections. In *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)* (pp. 444-451). IEEE.

[9] Bende, S., & Shedge, R. (2016). Dealing with Small Files Problem in Hadoop Distributed File System. *Procedia Computer Science*, *79*, 1001-1012.

[10] Cohen, J. C., & Acharya, S. (2014). Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection. *Journal of Information Security and Applications*, *19*(3), 224-244.