Available *online* at : http://jitce.fti.unand.ac.id/

JITCE (Journal of Information Technology and Computer Engineering)

| ISSN (Online) 2599-1663 |

Enhancing Approach for Information Security in Hadoop

Yogesh Awasthi [1], *Ashish Sharma* [2]

[1] *College of Engineering and Computer Science, Department of IT, Lebanese French University, Erbil, KR, Iraq*
[2] *College of Engineering and Computer Science, Department of Networking, Lebanese French University, Erbil, KR, Iraq*

ARTICLE INFORMATION

Received: December 8th, 2019
Revised: February 14th, 2020
Available online: March 30th , 2020

KEYWORDS

BigData, Hadoop, RSA, Rabin, Cryptography

CORRESPONDENCE

Phone: +964-7501618698
E-mail: dryogeshawasthi@lfu.edu.krd

ABSTRACT

Using the Hadoop package, a proposed secure cloud computing system was designed. Hadoop would use the area to establish and enhance the security for saving and managing user data. Apache had also produced a software tool termed Hadoop to overcome this big data problem which often uses MapReduce architecture to process vast amounts of data. Hadoop has no strategy to assure the security and privacy of the files stored in Hadoop distributed File system (HDFS). As an encryption scheme of the files stored in HDFS, an asymmetric key cryptosystem is advocated. Thus before saving data in HDFS, the proposed hybrid based on (RSA, and Rabin) cipher would encrypt the data. The user of the cloud might upload files in two ways, non-safe or secure upload files.

INTRODUCTION

Cloud computing has attracted increasing attention since the last few years. Cloud computing provides users with a wide range of resources such as computing platforms, storage, computing power, and internet applications. Amazon, Google, IBM, Microsoft, etc. are the biggest cloud available in the markets now. With a growing number of companies utilizing resources in the cloud, data from different users need to be protected. Cloud computing is presently being used in a tremendous amount in various fields. In daily life, huge amounts of data are produced. Consumers use cloud computing services to store this extremely large amount of data. Some of the major challenges cloud computing faces are to secure, protect and process the data that is the user's property [1]. Big data refers to the processing and retrieval of massive data collection. Big data must also be concerned with the collection of very important and sensitive data from social sites and issues of government and hence security. This collected data have to be encrypted by using appropriate algorithms to secure them. The characteristics of Big Data can be identified in term of four V's [2]: Volume, Velocity, Variety, and Veracity. Each issue has its very own activity of making due in Big information. Subsequently, volume: the measure of information delivered and may be put away it could be in the degree of numerous terabytes or Petabytes in size. Variety: which is the sort of information and its structures. Structure, semi structure and unstructured. Velocity: which indicated the input and the output rates of data streams generation and storing the system.in this context an abstraction is provided in a way that the enormous information frameworks can in the long run, store information freely of the approaching or active rate. Veracity: It's the term of data quality, this context is also Refers to the data confidentiality, data integrity privacy of data and availability of data.

Establishments must grantee that the data and the analysis conducted on the data are precise. Big data processing has become almost pivotal for many government and business applications with an incredible rate of data generated, collected and analyzed by computer systems. For Thus, many factors have participated in data huge increment like the emerge of IoT, object localization and tracking, besides the growing adoption of healthcare devices which gather personal statistics. This prevalence of big data, has some disadvantages. The data collected usually involves some personal information about persons or it is including secrets that would be problematic if they were discovered by the opponent. Criminal groups create underground markets for the possession and purchase of stolen personal information [3]. Government intelligence services rely on personal, corporate and adverse government eavesdropping and competitive ad- vantage systems. Most recent, highly publicized cyber-attacks against commercial attacks clearly demonstrate this potential for damage, and government targets, it

https://doi.org/10.25077/jitce.4.01.5-9.2020

pays millions of dollars to these organizations and causes severe damage to the affected individuals and organizations. Furthermore, security over cloud services is in its maturing phase, a large number of security vulnerabilities would risk data in the cloud. The cloud administrators have no clue as to where and in what format the data is being stored. The clients must be guaranteed in this situation that satisfactory safety efforts are to be adjusted to shield their data chiefly from information spillage and control of information. What's more, handling/examining colossal information at the cloud server farm is a basic issue. Distinctive circulated structures like HADOOP have as of late been accessible [4], similar to Google File System [5] which was created to store and process the Big Data. In any case, the dispersed HADOOP system is well known among industry and research networks. HADOOP includes two sets of functionalities, (i) For storage of large and unstructured data sets (HDFS), has been employed, and (ii) Map Reduce frame-work for hug data manipulation. HADOOP usually works with applications that have thousands of data nodes and petabytes.

HADOOP does not incorporate security mechanisms. The Application of ciphering algorithms in HADOOP data encryption then storing them at HDFS has been reported in several works. Ciphering schemes performs different replacements and do some manipulation on the clear message to transforms it into cipher text which must be random and unintelligible data. Different cipherment algorithms were developed and employed for sake of information security. Hence the two main categories are: (i) Symmetric-key cryptosystems [6] such as Data Encryption Standard (DES), Triple DES and Advanced Encryption Standard (AES) (ii) Asymmetric-key algorithms [7] like RSA and Elliptic Curve Diffe-Hellman (ECDH). The proposed approach can be considered as an attempt to improve what was presented by paper [8] at both of encipherment /decipherment of files of Big Data using Hadoop-integrated AES and OTP algorithms [8]. An architecture to secure Hadoop was examined in paper [9]. Thus for data encryption and decryption, AES encryption / decryption classes are added. Implement two HDFS-RSA and HDFS pairing integrations [10] used as some different types extensions of HDFS. Experiments exhibited an adequate overhead for understanding tasks and critical overhead for composing activities [11]. Three encryption scheme [12] were integrated in cloud data storage system depends on Hadoop to encrypt HDFS files based on RSA and DES then refer to IDEA to secure the RSA private key for the users, the encryption of the HDFS files is implemented when they are put away in a support in the wake of transferring information to HDFS. In this work a modified asymmetric key cryptosystem is being presented to secure Big Data. The following is the organization of this paper: Section II outlines the security framework. Section III, based on HDFS and MapReduce, presents the Big Data at HADOOP. Section IV discusses the proposed optimized hybrid encipherment algorithm and compare it with the classical public key cryptosystems prior to apply it to secure Big Data at HADOOP. Section V presents the discussion of the simulation results, finally, section VI list the conclusions.

SECURITY ISSUES

Big data is about data storage, data processing, data recovery. Many technologies, such as memory management, transaction management, visualization and networking, are used for these purposes. These technologies security issues are therefore also applicable to big data. Big data's four major security issues are authentication, data level, network level and generic issues [13].

Authentication Level Issues

With concentration on functional, non-functional requirements, system models, evolution of the system, and glossary of terms used in the project. A lot of clusters and nodes are present. Each node has priorities or rights that are different. Administrative nodes can access any data. But sometimes it will steal or manipulate the critical user data if any malicious node has administrative priority. Many nodes are joining clusters for faster execution with parallel processing. Any malicious node can disturb the cluster in the event of no authentication. Logging in big data plays an important role. If logging is not provided, no activity that modifies or deletes data will be recorded. If the new node joins the cluster, the absence of logging will not recognize it.

Data Level Issues

Data is a most important part of big data and also plays a vital role. Data is nothing but some of the government or social networking sites important and personal information about us. The main issues that could be handled by the Data level are integrity and availability of data like protection and distribution of data. Big data environments such as Hadoop store the data as it is without encryption to improve efficiency. If the hacker accesses the machines, it cannot be stopped. Information is stored in distributed data store for quick access in many nodes with replicas. But if hacker deletes or manipulates any replica or information from another node then it will be difficult to recover that data.

Network Level Issues

There are many nodes in clusters and these nodes are used to compute or process data. This data processing can be done anywhere between the cluster nodes. It is therefore difficult to determine which node data is being processed. It will be complicated because of this difficulty on which node safety should be provided. Two or more nodes can communicate or share their data / resources via network. RPC (Remote Procedure Call) is often used for network communication. But until and unless it is encrypted, RPC will not be secure.

General Level Issues

Many technologies are also used in the big data environment to process the data for some traditional security tools for security purposes. Over the years, traditional tools have been developed. Thus with new distributed form of big data these tools may not be performed well. As big data uses many data storage, data processing and data recovery technologies, there may be some complexities due to these different technologies.

BIG DATA AND HADOOP

Hadoop architecture consists mainly of two basic components which are :(HDFS) to store Big Data and MapReduce to analyze Big Data [14]. (HDFS) is a file management system used for the distributed storage of huge datasets on the Hadoop cluster in with

https://doi.org/10.25077/jitce.4.01.5-9.2020

a default block of 64 MB [15]. After storing the input files in HDFS then it calculated using a program identified by MapReduce eventually the results are moved to the output folder of HDFS [16]. Hadoop MapReduce is a software model designed to process large volumes of data sets over machine set [17]. MapReduce is the core scheme used by the Hadoop engine to distribute a cluster of work. The input file, which inhabit throughout the cluster on a distributed file system, is divided into groups of equal size to facilitate and simplify in a suitable, and almost error free manner the enormous amounts of data processing in parallel on large clusters of hardware. As specified by the name, MapReduce includes two – phases of information calculation in Hadoop, the principal stage is map and the subsequent stage is decrease, for example a huge volume of informational collections is changed over into organized key-esteem matches and given as data sources. MapReduce work [16].

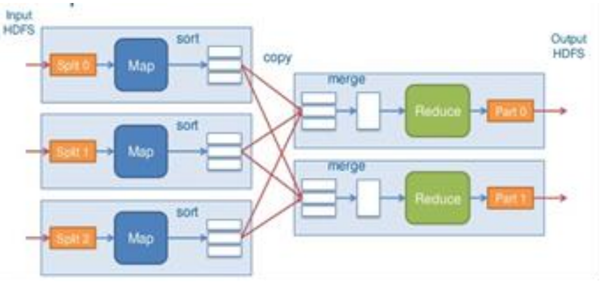


Figure 1. Data Flow of MapReduce computation

Fig. 1 shows the MapReduce calculation information stream. The mapper doesn't compose legitimately to the memory block yet exploits buffering the compositions. Every mapper has a round memory cushion with a default 100 MB size that can be altered by changing the property of (io. sort. mb). It makes an exceptionally astute flush. At the point when the cushion is topped off to a specific limit it started that spills the substance of the support to the circle. Before the spill happens to the circle, the string allotments the information as per the reducers it needs to go foundation string plays out a kind of in-memory inside the key-based parcel before the spill happens to the block. On the off chance that a combiner is available, it expends the yield of the in – memory sort [17].

PROPOSED ALGORITHM

There are several techniques available in cryptography, these techniques use methods of encryption to keep data from intruders secure. As Hadoop is the main provider of large-scale cloud data processing and storage, data security is a major concern. Hadoop therefore uses some techniques of encryption to ensure security. It uses the algorithm of AES encryption to encrypt data at rest. No encryption algorithm provides complete security, as mentioned above, and they all have their own limitations and loopholes that cannot be ignored. This paper introduces a new technique. This technique is based on two encryption algorithms' concept of hybridization. The limitations of encryption algorithms are overcome by this hybrid technique and security is improved. It's considered that all the files written to HDFS must be previously encrypted. Utilize Hadoop's new encryption system to encode the record while buffering to HDFS and utilizing unstructured information from the document. The HDFS starts working with the scrambled record in the wake of encoding the whole document. These stages have been shown in" Fig. 2." that's all. HDFS comprises of a Name Node that stores Metadata that deals with the namespace of the record framework and controls customers access to the scrambled document. The encoded document is comprised of at least one squares put away in a lot of information hubs.

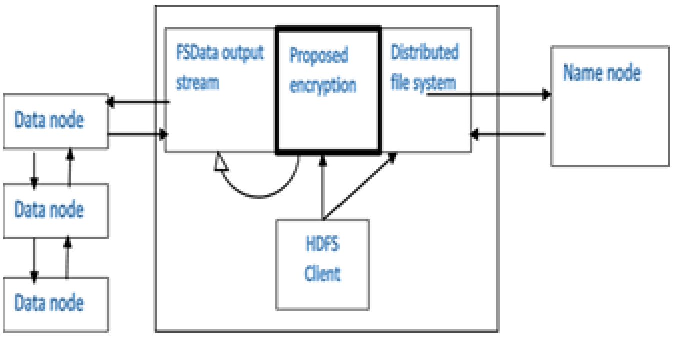


Figure 2. Encrypted Files in HDFS.

This method is hybrid in both algorithms between RSA and Rabin cryptosystem, using all key generation, encryption and decryption processes. The key generation process consists of both scenarios (RSA, Rabin) and two large prime (p, q) with the same size and their own multiplication calculates n: n= p*q, then Euler phi function is given by: Φ=(p-1) *(q-1), now chose random integer (e) such that **Error!** , and compute: $d = e^{-1} mod(n)$ (*Algorithm 1*) the public key for any the public key for any side will be (n,e) and the private key (p,q,d). p,q from Rabin , (n,e) and d from RSA. The encryption process is hybrid between RSA encryption and Rabin encryption to encrypt a message(m) the cipher text is computed as: $c = m^{2e} mod(n)$ (*Algorithm 2*) While the decryption process takes the same scenario of both RSA and Rabin. To recover the plain text from cipher text we use $m^2 = c^d mod(n)$ then compute the original message by computing the of m^2 which might be one of the m_1, m_2, m_3, m_4 messages explained in (*Algorithm3*). The encryption scheme of First Proposed Method is: On instruction from the Name Node, the Data Nodes are responsible for block creation, deletion, and replication. Client information is taken from different sources and scrambled on the server utilizing asymmetric mechanisms instruments of cryptography. Not long after encoded, information is put away in the cloud, i.e. via HADOOP File System (HDFS)

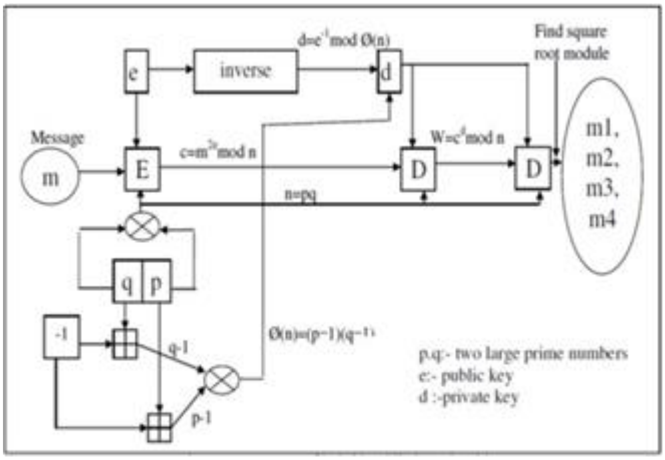
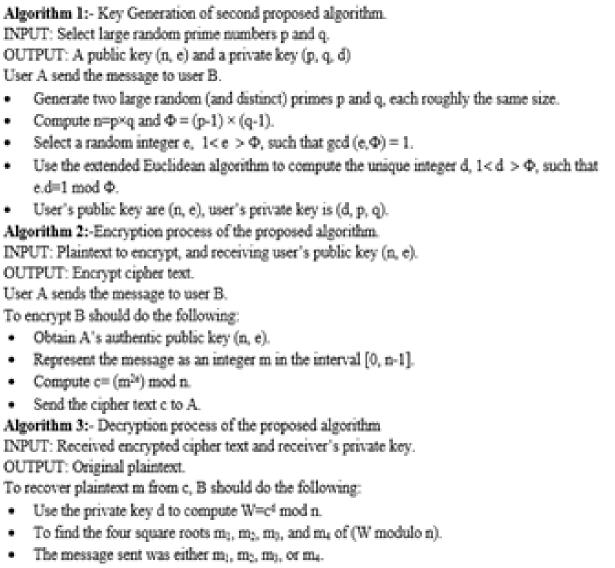


Figure 3. Process of Hybrid Key Algorithm.

it will be stored in a cluster. Whenever the user requests data, the server will provide the encrypted data for decryption. The user then uses the corresponding keys to retrieve the decrypted data using a hybrid approach which is suggested in this paper. There are several steps of data encryption as shown in Fig. 3.

EXPERMENTAL RESULTS AND ANALYSIS

Hadoop 2.7.1 has been arranged as solitary hub bunch to utilize the HDFS and Mapreduce capacities for execution assessment of encoded HDFS. Each node has i3 core, 4 processors, 4 GB of memory, and 750 G of hard disk. Two different methods of encryption / decryption have been employed, by using RSA algorithm the time of record encryption between nonexclusive HDFS and encoded HDFS using Hybrid algorithm to perform encryption / decryption was compared with other classical algorithms. Fig. 4, depicts the results of the comparison between encryption schemes the RSA and the Hybrid asymmetric encryption algorithms with different file sizes. It's clear from Fig .4 that the proposed method showed efficient time consumption compared to the RSA, with comparable complexity to RSA cryptosystem. For files size stars from 100 MB and ends with 1 GB with step size of 100 MB.



Algorithm 1:- Key Generation of second proposed algorithm.
INPUT: Select large random prime numbers p and q.
OUTPUT: A public key (n, e) and a private key (p, q, d)
User A send the message to user B.

- Generate two large random (and distinct) primes p and q, each roughly the same size.
- Compute $n=p\times q$ and $\Phi = (p-1) \times (q-1)$.
- Select a random integer e, $1< e > \Phi$, such that $\gcd(e,\Phi) = 1$.
- Use the extended Euclidean algorithm to compute the unique integer d, $1< d > \Phi$, such that $e.d=1 \bmod \Phi$.
- User's public key are (n, e), user's private key is (d, p, q).

Algorithm 2:-Encryption process of the proposed algorithm.
INPUT: Plaintext to encrypt, and receiving user's public key (n, e).
OUTPUT: Encrypt cipher text.
User A sends the message to user B.
To encrypt B should do the following:

- Obtain A's authentic public key (n, e).
- Represent the message as an integer m in the interval [0, n-1].
- Compute $c= (m^{2e}) \bmod n$.
- Send the cipher text c to A.

Algorithm 3:- Decryption process of the proposed algorithm
INPUT: Received encrypted cipher text and receiver's private key.
OUTPUT: Original plaintext.
To recover plaintext m from c, B should do the following:

- Use the private key d to compute $W=c^d \bmod n$.
- To find the four square roots m_1, m_2, m_3, and m_4 of (W modulo n).
- The message sent was either m_1, m_2, m_3, or m_4.

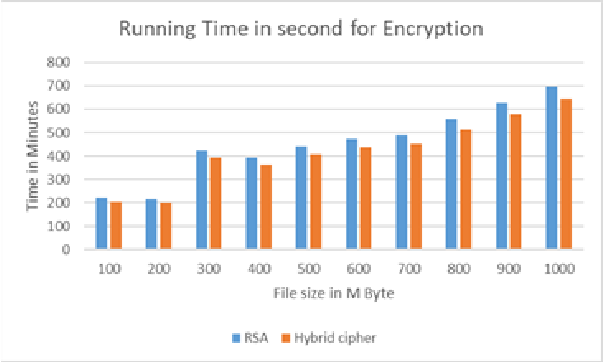


Figure 4. Running Time in second for Encryption schemes

Thus, the proposed method (Hybrid cipher) in encryption stage is faster than the default RSA. Fig .5 shows the running time for RSA and proposed method in decryption stage. The encrypted files applied to this stage are with different sizes. By applying both of RSA and the hybrid cipher (the proposed method) it's obvious that decryption time needed by hybrid ciphered method is shorter than that needed by RSA. Table 1. Shows the computational complexity of the Hybrid cipher (proposed method), RSA and Rabin, from which it's clear that the proposed method has more complex than the RSA and Rabin.

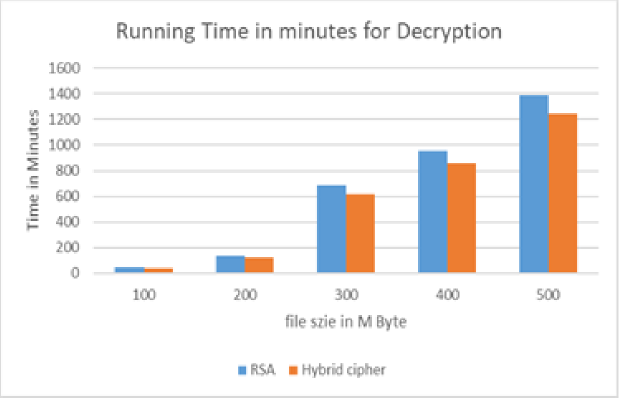
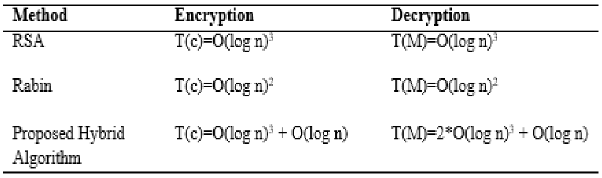


Figure 5. Running Time in minutes for Decryption schemes

Table 1. Computational Complexity of the Proposed Hybrid Method



Method	Encryption	Decryption
RSA	$T(c)=O(\log n)^3$	$T(M)=O(\log n)^3$
Rabin	$T(c)=O(\log n)^2$	$T(M)=O(\log n)^2$
Proposed Hybrid Algorithm	$T(c)=O(\log n)^3 + O(\log n)$	$T(M)=2*O(\log n)^3 + O(\log n)$

CONCLUSIONS

While Hadoop allows us to overcome the challenges faced by big data in industries and institutions, it has no security mechanism. An attacker or eavesdropper may compromise the data stored in Hadoop. The authenticity of data is always at stake, as Hadoop does not provide any security mechanism. Before storing it in HDFS, the proposed Hybrid cipher asymmetric key algorithm encrypts the file content by securing it from the various network attacks. The information or documents can in this way currently be put away in Hadoop without agonizing over security issues by applying the encryption calculation to the records before it is put away in Hadoop. The proposed system supports most cloud computing system service models such as Service Software (SaaS), Service Infrastructure (IaaS), and Service Platform (PaaS). It also supports data management and security issues (Authentication, Integrity, Availability, and Confidentiality) in security and key management for data transfer. The proposed method shows excellent time consumption with different file sizes in the encryption and decryption stage with higher complexity.

REFERENCES

[1] P. Merla and Y. Liang, "Data analysis using Hadoop MapReduce environment," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 4783-4785.doi: 10.1109/BigData.2017.8258541.

[2] Martin Hilbert, "Big Data for Development: A Review of Promises and Challenges", Journal Development Policy Review of the Overseas Development In- stitute,2016.

[3] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, M.Voelker. An analysis of underground forums. In Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, IMC '11, pages 71{80, NewYork, NY, USA, 2011. ACM.

[4] S. Park, Y. Lee, "Secure Hadoop with Encrypted HDFS," Chapter Grid and Per- vasiveComputing, Vol. 7861 of the series Lecture Notes in Computer Science, pp 134-141, (2013).

[5] Bo Li, Mengdi Wang, Yongxin Zhao, Geguang Pu, Huibiao Zhu, Fu Song " Modeling and Verifying Google File System Modeling and Verifying Google File System" , 16th International Symposium on High Assurance Systems Engineer- ing, Pages: 207 -214,IEEE,(2015).

[6] Sourabh Chandra, Siddhartha B, Smita Paira." A Study and Analysis on Symmetric Cryptography" ICSEMR, pp 1-8, IEEE (2014).

[7] Sourabh Chandra, Sk Safikul Alam, Smita Paira and Goutam Sanyal. "A comparative survey of symmetric and asymmetric key cryptography", International Conference on Electronics, Communication and Computational Engineering (ICECCE), pp 83-93,IEEE (2014).

[8] Hadeer Mahmoud, Abdelfatah Hegazy, and Mohamed H. Khafagy "An approach for Big Data Security based on Hadoop Distributed File system", International Conference on Innovative Trends in Computer Engineering (ITCE 2018), Aswan University, Egypt, 2018.

[9] Park and Y. Lee, "Secure Hadoop with Encrypted HDFS," pp. 134–141, 2013.

[10] H. Y. Lin, S. T. Shen, W. G. Tzeng, and B. S. P. Lin, "Toward data confidentiality via integrating hybrid encryption schemes and Hadoop distributed file system," Proc. - Int. Conf. Adv. Inf. New. Appl. AINA, pp. 740–747, 2012.

[11] M. M. Shetty and D. H. Manjaiah, "Data security in Hadoop distributed file system," Proc. IEEE Int. Conf. Emerg. Technol. Trends Computer. Commun. Electr. Eng. ICETT 2016, pp. 939–944, 2017.

[12] C. Yang, W. Lin, and M. Liu, "A novel triple encryption scheme for Hadoop- based cloud data security," Proc. - 4th Int. Conf. Emerg. Intell. Data Web Technol. EIDWT 2013, pp. 437–442, 2013.

[13] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri. seccurity Issues Associated With big Data In Cloud Computing.

[14] Bhandarkar, M, "MapReduce programming with apache Hadoop," IEEE Inter- national Symposium on Parallel & Distributed Processing (IPDPS), pp.1-2, 2010.

[15] Aditya Bhardwaj, Vineet Kumar Singh, Vanraj, Yogendra Narayan, "Analyzing BigData with Hadoop Cluster in HDInsight Azure Cloud", Annual IEEE India Conference (INDICON), 2015.

[16] Dubey, A. K., Jain, V., & Mittal, A. P," Stock Market Prediction using Hadoop Map-Reduce Ecosystem" IEEE 2nd International Conference on Computing for Sustainable Global Development, pp.616-621, 2015.

[17] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In Communications of the ACM, 51 (1): 107-113, 2000.