www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-5, Issue-4, August 2017
Research Paper

Int. J. Sc. Res. in Network Security and Communication

ISSN: 2321-3256

Providing Security to Data Stored on HDFS Using Security Protocol

Poonam R. Wagh[1*], Amol D. Potgantwar[2]

[1*]Dept. of Computer Engineering, Sandip Institute of Technology and Research Centre, Nashik, India
[2]Dept. of Computer Engineering, Sandip Institute of Technology and Research, entre, Nashik, India

**Corresponding Author: wagh.poonam@gmail.com*

Received 13th Jun 2017, Revised 27th Jun 2017, Accepted 18th Jul 2017, Online 30th Aug 2017

Abstract—Different organizations and individuals tend to outsource their data to cloud storage, the security and user privacy protection attract more attention. Several attempts has been made to improve security of the cloud storage, which used for security of data but it is not very effective method for protecting and securing confidential data from unauthorized access. In the proposed system combination of security systems used to ensure security of the network to additional extent. A novel model of cloud secure storage is proposed, which combines the Hadoop distributed file system (HDFS) security protocol and cryptography for data stored on HDFS. The model uses the HDFS as the storage platform. Hadoop uses security protocols like Kerberos, LDAP and keytabs for user authentication. The fast encryption of cryptography algorithms and identity authentication like RSA, blowfish, AES can be used for overtime checking and the performance of Hadoop. Thus the security system with combination at different levels can supply secured, effective, stable effect on large data through Hadoop and cloud computing.

Index Terms—HDFS, Kerberos, Cloud computing, AES

I. INTRODUCTION

Nowadays data which is being collected and processed is extremely large, big data is extremely large data sets that are analyzed and converted into different patterns, trends, and associations. There are three types in big data structured, semi structured and unstructured. Big data technologies are important in providing accurate analysis. Big data faces many challenges like capturing data, duration storage, searching, sharing, analysis and security. Big data analytics is often associated with cloud computing because the analysis of large data sets in real-time needs a platform like Hadoop to store large data sets across a distributed cluster and Map Reduce to coordinate, combine and process data from multiple sources. Cloud computing is extremely popular and highly used com- puting paradigm, it provides the users massive computing, storage, and software resources on demand [1]. Hadoop is a Distributed framework for analyzing big data [2].It is a platform for structuring Big Data, which solves the problem of formatting it for subsequent analytics purposes. Hadoop has a distributed computing architecture with multiple servers, mak- ing it extremely inexpensive to scale and support extremely large data stores. DFS mainly consists of Name Node, Data Node, Job tracker and task tracker [3]. Owen OMalley etal. designed the Kerberos protocol based on SSL to launch user identity authentication [4]; Indrajit Roy et al. designed and implemented the Airvat platform, which could ensure the vital data secure and privacy protection in the Map Reduce calculation process[5].Hadoop distributed file system is based on the Map Reduce processing technique. Map Reduce is a processing technique model for distributed computing based on java [6]. There are two important tasks in it, namely Map and Reduce. Map takes a set of data and converts that into another set of data, the files are split into the lines. Reduce task, takes the output from a map as an input and combines those data rows into a smaller set of rows. As the sequence of the name Map Reduce, the reduce task is always performed after the map job. Map Reduce has a major advantage that it is easy to scale data processing over multiple computing nodes. These are data processing primitives, called Mappers and reducers. The core concept is processing speed across large data sets. Breaking large data sets into small pieces, distributing them to as many storage/processing units as possible, and processing data, such that processing and data are tightly coupled with the resulting output being aggregated are the key feature to achieving the goal of speed. A framework of SecureMR has been proposed by Wei Wei et al. to guarantee the data and service integrity[7]. In fact, the research works they have done were mostly aimed at providing the users authentication to identify and to ensure the security as well as privacy protection. In Hadoop Distributed File systems (HDFS) there is no attempt to verify the identity and group membership of users who interact with (HDFS) and logged in users store data in the Hadoop in a browser without any additional storage media, and the user can obtain the data wherever they are by computers, laptops, mobile phones. It

is not having effective security for the confidential data which is not access to unauthorized user. Kerberos protocol messages are protected against eaves dropping and replay attacks. It builds on symmetric key cryptography and requires a trusted third party, and optionally may use public-key cryptography during some phases of authentication [7]. In Hadoop, Kerberos authentication is used but authentication is not very effective for securing confidential data on Hadoop [8]. Below are the current issues on the HDFS Security 1) Transmission security: While transmitting data it may be intercepted, but the data transmission is not working with the strong encryption protec- tion measures. 2) Access control: Authorized access control is weak the user data stored in the clouds without setting access authority. 3) Data storage: Data stored on the cloud is not classified hence data leakage is possible 4) Data verification: Data verification is not strong as it cannot be verified if the right person is accessing the data. To solve the existing security problems, cloud disk storage based on Hadoop is proposed, the program draw lessons from a security protocols like Kerberos authorization process. The classic algorithms such as AES, RSA, blowfish for realizing encryption, and authentication are utilized, and time is checked to inspect if it can complete the encryption and transmission in an acceptable period.

II. REVIEW OF LITERATURE

Network essentials and applications are proposed by Stallings [9].Hadoop eco system for big data security is proposed by pradeep adluru[10]. David Nunez, Isaac Agudo proposed Cryptographically Enforced accessed control system, Encryption and decryption on data at job tracker which faces performance issue[11]. There is Implementation of SEHadoop model which has delegation token limited performance im- pact[12]. Certificate and timestamp are checked before lo- gin and TSA and directory to verify on server by secure cloud management method[13].There is a solution proposed by yan Wen for Hadoop but it faces security in terms of efficiency[14].Data at rest security is proposed for Hadoop security threats this system uses tuning for hdfs[15]. Auth hacker and perimeter eavesdropper is used by Sharif, Sarah Cooney[16].Spark uses HDFS security system and analysis for spark cloud computing security is done using static secu- rity analysis[17].One solution provided in combination with Kerberos for authentication and single sign on (SSO) to make the system secure [18].Many security and privacy issues are discussed by Ibrahim Lahmer, Ning Zhang[19]. The RSA digital signature algorithm and its mathematics are used and proved [20]. Large and integrated integer library is built by using C++ and the implementations of Miller-Rabin. Hong Bo Zhou explained cloud computing its applications and standards [21].There is Hatman intra cloud trust management system in which the first full-scale, reputation-based ,trust management system is implemented. By comparing job replica outputs for consistency, Hatman dynamically assesses node integrity [22].BANLOGIC is implemented by distributing encryption system to reduce burden on the system to achieve security and stability[23].XU Guang-hui, in his research mentioned, Hadoop as an Apache open source project consisting of several projects like HDFS, MapRedcue, HBase, Hive, ZooKeeper and others [24]. Further to his findings it was concluded that HDFS and MapReduce being vital parts of the open source. According to ZHUO Tang, in his findings aimed at Algorithm for MapReduce,MapReduce had aimed at paralleling and dealing with tasks on a much larger scale. Paralleling and Dis- tributed Processes had eventually made MapReduce scheduler become particularly more important [25]. While XU Guang- huis MapReduce scheduler became important in parallel and distributed processes of the open source, the HDFS(Hadoop Distributed File System) again an open source from Google distributed source (GFS)[26] a cumulative effort by Ghemawat S, Gobioff H, Leung S T. came to light. It came with its own pros and cons one being that it has a high fault tolerance and only a certain data access control. Another disadvantage of Hadoops HDFS was that it could mainly be used for cloud storage only, limiting its uses. Ghemawat also stated that Hadoop lacked serious safety measures making it vulnerable to data leakages. To overcome such vulnerabilities Ghemawat S. integrated Kerberos in Hadoop in the year 2009 with the help of Yahoo. Researchers of Kerberos suggested its implementation by only authorizing the access to users. According to the researchers users of Kerberos had to obtain access from third party centers which in turn will produce an authorization certificate only after which the users will be allowed access. After the authorized certificate, Hadoop Cluster will sort key issues first depending upon the issue minimizing the risk of user data theft. Many other researchers put their theories in different methods to reduce data leakages caused identification and to increase the security of cloud storage. [27]ZHANG Da-Wei, 2009, are searcher on Hadoop -based enterprise file cloud storage system proposed that the use of HDFS can be used to build a private enterprise cloud which can combine Hadoops fault tolerance and can also be suitable for the Big Data feature. In addition to building a secure and private enterprise, [28]HouQinhua,Wu Yongwei,Zheng Weimin, researched on Protectionof User Data Privacy in Cloud Storage Platform. Hou, further evaluated that a System Security Layer (SSL)secure connection and secure virtual machine monitor could enhance the security of user data in cloud storages. Furthermore, YU Shu-cheng[29], encoded the existing cloud data using the attribute encryption scheme (ABE), the work was simple but effective by simply using a public key to decode the same data before even it was uploaded, the use of public key limited the data access the user had. The user had to ensure to use the K attributes

to decrypt the data, in which the K is the number of threshold to decrypt the data. The use of the attribute scheme had ensured the safety and security of the data storage but eventually eliminated the use of public key for each individual user at the server side. YU Shucheng scheme of attributes enabled the user attributes to be used as users public key which also came along with some cons, the data could be decrypted completely when different user attributes would hold their attributes together and get access to all other attributes.

III. Software Requirement Specification

A software requirements specification is a comprehensive description of the intended purpose and environment for the proposed work under development. The SRS fully describes what the proposed work will do and how it will be expected to perform. System user will be any user who uses internet for retrieving information. Depending on the type of cloud provider the data access authority will be assigned to user. The use of your data may occur unbeknownst or by virtue of a configuration error on the providers part. Based on the sensitivity of the data, contract prohibits or at least limits the access the cloud provider has to use this data. The user should have basic knowledge of computer and English language for providing the information query. Then user should have basic knowledge of accessing the database. The user must be authenticated user. The user also knows how to access the data stored in cloud. The key used for encryption and decryption is kept secure from outsiders. The user should have basic knowledge about the cloud database. System will have database information which is stored in cloud. User access cloud computing using networked client devices, such as desktop computers, laptops, tablets and smartphones and any Ethernet enabled device such as Home Automation Gadgets. This is useful for the huge database industries to put their data over cloud and search data using multi keyword. The database is very big in size, also provides higher security than any other database. The large industries can use cloud computing for storing there database.

IV. System Overview

In proposed system we provide security to Hadoop to improve the data security and trusted condition of Hadoop cluster. All HDFS clients must be authenticated to ensure that the one who is authenticating is who he claims to be and providing the secure data storage for the user by adding security levels in HDFS. The block diagram above shows basic architecture of the Hadoop cluster.

It has mainly a name node which reads the metadata of incoming file and stores all the information about the file. It makes the copies of the incoming file based on the number of times mentioned in the inode. Job tracker and task tracker performs the related operations on the file. Data is stored on the data nodes. There is no security provided to the data on the HDFS system. Hence new security system is proposed. In which a highly secure protocol is implemented to embraced combination of Kerberos five and AES cryptography which will enhance security in cloud computing.

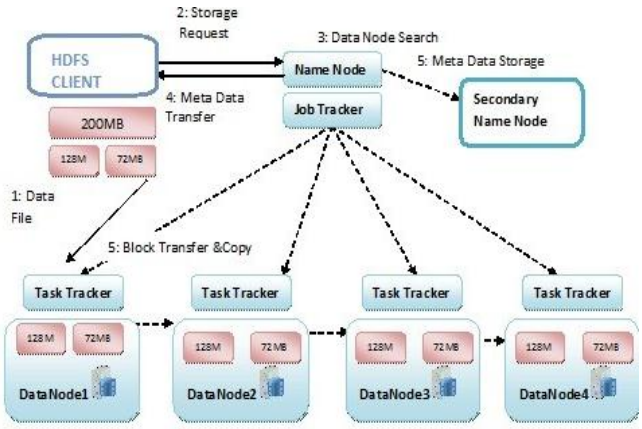

Fig. 1. Hadoop Cluster

- Kerberos Authentication Service

Kerberos is one amongst secure technique for authenticating letter of invitation for a service in a network. Kerberos was developed within the Pallas at the Massachusetts Institute of Technology. Kerberos assigns tickets to Kerberos principals to enable them to access Kerberos-secured Hadoop services. A Kerberos principal is used in a Kerberos secured system to represent a unique identity.

- AES Algorithm

AES algorithm was revealed by bureau in 2001. It is a bilaterally symmetric Block cipher with a block length of 128 bits and support for key lengths of 128,192 and 256 bits. Bureau handpicked Rijindael because the projected AES formula. Dr. Joan Daemen and Dr. Vincent Rijmen are two researchers who developed and submitted Rindael for the AES. Decryption in AES algorithm involves reversing all the steps taken in secret writing mistreatment inverse functions like InvSubBytes, InvShiftRows, and InvMixColumns.

Using Kerberos as security authentication system and AES as distributed cryptography algorithm we propose a secure and time efficient HDFS cloud system. A security system from the client to server cloud is proposed as shown in Fig.2. The authentication of the client is done to check if the right person is accessing the data.

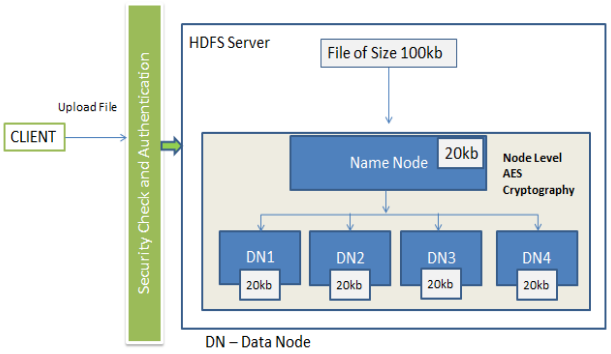

Fig. 2. Distributed Cryptography with Security Protocol

Then security protocol is implemented and data is stored by using the encryption algorithm. While retrieving the data the respective decryption algorithm is used and data is displayed as per the authority of the client. The encryption and decryption is done at node level which processes file parallel there by reducing the time required for cryptography. In proposed system the security of the cloud is assured by proving the security on different layers in the cloud. Security protocol will be the combination of security protocol.

V. System Analysis

A mathematical model is an explanation of a system using mathematical concepts and words. The procedure of developing a mathematical model is termed mathematical modeling. Fig 3 shows the mathematical model of system. The Mathematical explanation of the proposed system is explained as, U1 is client for Authentication in which user login if user login successfully then proceed for further process. user may be sender(U1) or receiver(U1) where,

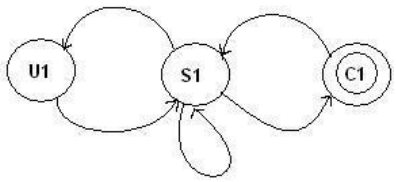

Fig.3. Proposed System State diagram

Cn is Cloud Data storage. After that U1 is function of sender it gives the permission to data for security and proceeds to next function. HDFS security i.e. S1 function encrypt data. f22 function uploads the data on clouds.

The parameters are explained in following section.
Input (I) Parameter: I= [Dn] Where, I is a set of Input. Dn= Encrypted Data or Plain text. Function(F) Parameter F=[U1,S1,C1] Where, F is a function for processing. U1= Client User. S2= Security Model. C3= Data on cloud. Output (O) Parameter O=[f1] Where, O is the Output. f1=Encrypted or original data. Process of the model is:

1.Request to connect cloud. 2. Encrypted text and client details with key. 3. Request to connect client authentication. 4. Send a file with encrypted text and key for algorithm. 5. Request to fetch file from HDFS cloud with key. 6. Verify client and retrieve encrypted file. 7. Encrypted file sent to client.

VI. Result Analysis And Comparison

We aim at the existing popular cloud security weakness. Therefore as per our proposed system, encryption and decryption is performed at the HDFS cloud using distributed cryptography technique. The files will be encrypted and decrypted at server side hence performance measure will not depend on the local system. The distributed cryptography performs depending on the number of nodes in HDFS cloud. We performed test on different number of nodes using data set of 20 Newsgroups. Following table shows the results of the experiment.

Table 1: Result Analysis

No. of Nodes	With Encryption Time(sec)	Without Encryption Time(sec)
1	427.55	352.36
3	368.67	326.45
5	321.12	286.64

Below graph shows the encrypted data storage on cloud. Time required encrypting and decrypting data compared to the existing system.

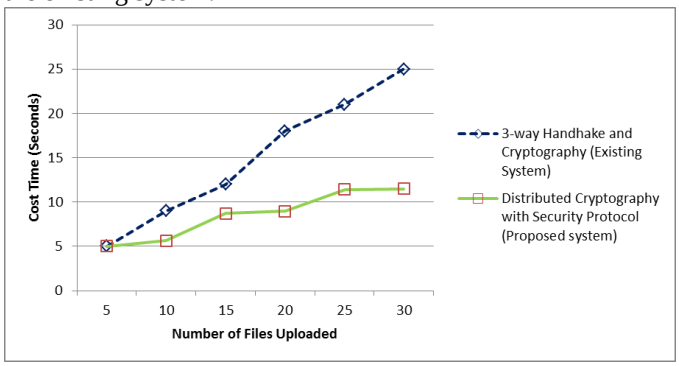

Fig. 4.Distributed Encryption using AES Algorithm

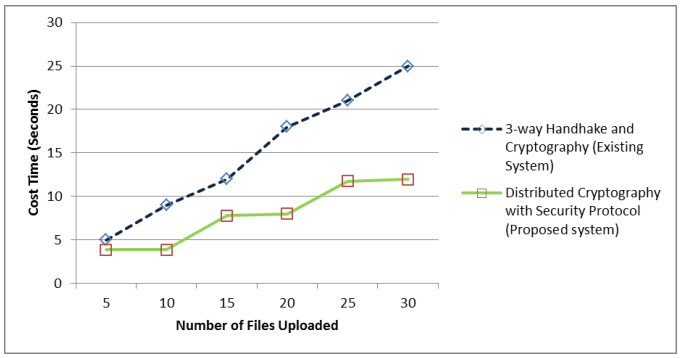

Fig. 5.Distributed Decryption using AES Algorithm

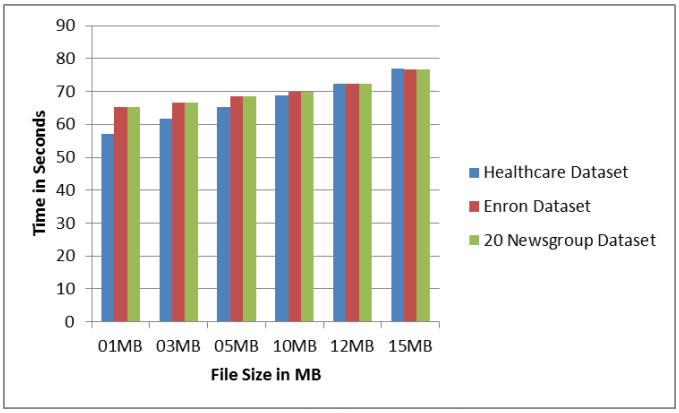

Fig 6.Comparision of different types of datasets

VII. Comparison With Similar Systems

Kerberos supports symmetric as well as asymmetric cryptography unlike NTLM. LDAP can easily misconfigure to send credentials in clear text over the network, encryption is not used. LDAP encapsulates all traffic in SSL. LDAP is often used for adhoc authentication/authorization especially web applications using forms authentication. Kerberos is used for user authentication compared with LDAP which is used for user authorization. Compared with LDAP, Kerberos is preferred because it is more secure. Distributed AES cryptography is parallel processing of the file which is a scalable solution as we increase the number of nodes the performance will increase.

VIII. Conclusion

The security to HDFS cloud is provided by using the security protocol and cryptography. Data is stored on the cloud in encrypted format which can only be decrypted by the client. Hence the security to the data stored on the cloud will be improved and made more secure.

Acknowledgment

This work is supported by SITRC(Sandip Institute of Technology and Research center), Nasik, Maharashtra, under the guidance of respected Dr. Amol D. Potgantwar, Head of Department Computer Engineering. This work would not have been possible without the enthusiastic response, insight, and new ideas from him.

References

[1]. Nitesh Jain, Pradeep Sharma, "*A Security Key Management Model for Cloud Environment*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.1, pp.45-48, 2017.

[2]. B. Subasri, P. Vijayalakshmi, P. Yurega, E. Revathi, "*Improving Zero Knowledge in Cloud Storage Auditing System*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.3, pp.204-207, 2014.

[3]. Rajesh Verma , "*Comparative Based Study of Scheduling Algorithms for Resource Management in Cloud Computing Environment*", International Journal of Scientific Research in Computer Science and Engineering, Vol.1, Issue.4, pp.17-23, 2013.

[4]. B. Lakhe, "*Monitoring in Hadoop*", Practical Hadoop Security, pp. 119–141, 2014.

[5]. Shubhangi D.C., Sabahat Fatima, "*Privacy-Preserving Outsourcing of Medical Image Data using SIFT Descriptor*", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.3, pp.141-145, 2017.

[6]. Y. Wang, C. Xu, X. Li, and W. Yu, "*JVM-Bypass for Efficient Hadoop Shuffling*," 2013 IEEE 27th International Symposium on Parallel and Distributed Processing, May 2013.

[7]. Ruchi Mittal and Ruhi Bagga, "*Performance Analysis of Hadoop with Pseudo-Distributed Mode on Different Machines*", International Journal of Computer Sciences and Engineering, Vol.3, Issue.6, pp.113-117, 2015.

[8]. J. Cohen and S. Acharya, "*Towards a Trusted Hadoop Storage Platform: Design Considerations of an AES Based Encryption Scheme with TPM Rooted Key Protections*," 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing, Dec. 2013.

[9]. W. Stallings, "*Network Security Essentials: Applications and Standards*", Prentice-Hall Published, 2000.

[10]. P. Adluru, S. S. Datla, and X. Zhang, "*Hadoop eco system for big data security and privacy*," 2015 Long Island Systems, Applications and Technology, May 2015.

[11]. D. Nunez, I. Agudo, and J. Lopez, "*Delegated Access for Hadoop Clusters in the Cloud*," 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, Dec. 2014.a

[12]. Xianqing Yu, P. Ning, and M. A. Vouk, "*Enhancing security of Hadoop in a public cloud*," 2015 6th International Conference on Information and Communication Systems (ICICS), Apr. 2015.

[13]. Lin Wen-hui, Lei Zhen-Ming, Yang Jie, He Gang, Liu Jun, and Liu Fang, "*Secure cloud storage management method based on time stamp authority*," National Doctoral Academic Forum on Information and Communications Technology 2013, 2013.

[14]. H. Zhou and Q. Wen, "*A new solution of data security accessing for Hadoop based on CP-ABE*," 2014 IEEE 5th International Conference on Software Engineering and Service Science, Jun. 2014.

[15]. P. Zerfos, H. Yeo, B. D. Paulovicks, and V. Sheinin, "*SDFS: Secure distributed file system for data-at-rest security for Hadoop-as-a-service*," 2015 IEEE International Conference on Big Data (Big Data), Oct. 2015.

[16]. A. Sharif, S. Cooney, S. Gong, and D. Vitek, "*Current security threats and prevention measures relating to cloud services, Hadoop concurrent processing, and big data*," 2015 IEEE International Conference on Big Data (Big Data), Oct. 2015.

[17]. G. Zhou, D. Zhao, K. Zou, W. Xu, X. Lv, Q. Wang, and W. Yin, "*The static security analysis in power system based on Spark Cloud Computing platform*," 2015 IEEE Innovative Smart Grid Technologies - Asia (ISGT ASIA), Nov. 2015.

[18]. A. Desai, Nagegowda K S, and Ninikrishna T, "*Secure and QoS aware architecture for cloud using software defined networks and*

Hadoop," 2015 International Conference on Computing and Network Communications (CoCoNet), Dec. 2015.

[19]. Arsha Sultana and S. Madhavi , "*Generating Optimized Association Rule for Big Data Using GA and MLMS*", International Journal of Computer Sciences and Engineering, Vol.3, Issue.9, pp.144-148, 2015.

[20]. V. Kapoor, "*Data Encryption and Decryption Using Modified RSA Cryptography Based on Multiple Public Keys and 'n'prime Number*", International Journal of Scientific Research in Network Security and Communication, Vol.1, Issue.2, pp.35-38, 2013.

[21]. S.L.Mewada, U.K. Singh, P. Sharma, "*Security Enhancement in Cloud Computing (CC)*", International Journal of Scientific Research in Computer Science and Engineering, Vol.1, Issue.1, pp.31-37, 2013.

[22]. S. M. Khan and K. W. Hamlen, "*Hatman: Intra-cloud Trust Management for Hadoop,*" 2012 IEEE Fifth International Conference on Cloud Computing, Jun. 2012.

[23]. F. A. H. Jing, S. B. L. Renfa, and T. C. T. Zhuo, "*The research of the data security for cloud disk based on the Hadoop framework*," 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), Jun. 2013.

[24]. XU Guang-hui, *"Deploying and researching Hadoop in virtual machines (ICAL)"*, IEEE International Conference on Zhengzhou, pp. 395-399, 2012.

[25]. Z. Tang, J. Zhou, K. Li, and R. Li, "*MTSD: A Task Scheduling Algorithm for MapReduce Base on Deadline Constraints*," 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum, May 2012.

[26]. P.N. Priyanka, S.V. Phaneendra, "*Study on Migration from conventional File System to advancement of Bigdata Technologies in the real world*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.8, pp.11-20, 2014.

[27]. Da-Wei Zhang, Fu-Quan Sun, Xu Cheng, and Chao Liu, "*Research on hadoop-based enterprise file cloud storage system,*" 2011 3rd International Conference on Awareness Science and Technology (iCAST), Sep. 2011.

[28]. Hou Qinhua,Wu Yongwei,Zheng Weimin, "*A Method on Protection of User Data Privacy in Cloud Storage Platform*", Journal of Computer Research and Development, pp.4871146-1154, 2011.

[29]. S. Yu, C. Wang, K. Ren, and W. Lou, "*Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing,"* 2010 Proceedings IEEE INFOCOM, Mar. 2010.

Authors Profile

Ms Poonam R. Wagh has pursued Bachelor of Engineering in Computer from University of Pune in 2008. She is now pursuing Master of Computer Engineering from Savitribai Phule Pune University.

Dr. Amol D. Potgantwar is an Associate Professor of the Department of Computer Engineering, Sandip Foundation's, Sandip Institute of Technology and Research Centre, Nashik, Maharastra, India. The focus of his research in the last decade has been to explore problems at Near Field Communication and it's various application In particular, he is interested in applications of Mobile computing, wireless technology, near field communication, Image Processing and Parallel Computing. He has register patents like Indoor Localization System for Mobile Device. Using RFID & Wireless Technology , RFID Based Vehicle Identification System And Access Control Into Parking, A Standalone RFID And NFC Based Healthcare System. He has recently completed a book entitled Artificial Intelligence, Operating System, Intelligent System. He has been an active scientific collaborator with ESDS, Carrot Technology, Techno vision and Research Lab including NVIDIA CUDA, USA. He is a member of CSI, ISTE, IACSIT.

