ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

Hadoop MapReduce Authentication and Privacy

Prajwal.M.H , Puneeth.H.V , Sukruth.R , Chetan Kashyap[1] , Raghavendra Babu .T.M[2]

B.E, Students, Dept. of CS&E, P.E.S.C.E., Mandya, Karnataka, India [1]

Asst. Prof. Dept. of CS&E, P.E.S.C.E., Mandya, Karnataka, India [2]

ABSTRACT: - Hadoop is an open source framework written in Java. It supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project. MapReduce is a Hadoop framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. Data processing is carried out in two phases: map and reduce. The map phase takes a set of data and converts it into another set of data called key/value pairs to produce the intermediate results of the MR computation. The reduce phase then takes these intermediate results as its input and combines these data to produce an output and this output is the final result of the MR computation. A major concern of using the MR model in a public cloud is its in adequate security provision, such as authentication. The MR model was initially intended for use in private networks, so the issue of security was not a design consideration. Since its introduction, lots of efforts have been made to improve the performance of this model making it more efficient rather than making it more secure. Deploying the MR model in an open environment, such as public clouds, without adequate security provisioning would put the clients' jobs and their data at risks. This is because, in such an environment, different jobs submitted by different clients typically share the same set of physical nodes and software resources which poses a potential risk to the client's security.

KEYWORDS: Hadoop MapReduce, Authentication, AES Encryption, Privacy.

I. INTRODUCTION

Hadoop MapReduce framework is a new parallel programming paradigm which is proposed to process large volumes of data. Data processing is carried out in two phases: map and reduce The guide stage takes an arrangement of information and believers it into another arrangement of information called key/esteem sets to deliver the moderate after-effects of the MR calculation . The decrease stage at that point takes these middle of the road comes about as its info and consolidates these information to deliver a yield and this yield is the final after-effect of the MR calculation. To do the two-stage MR calculation, an arrangement of appropriated hubs (from this point forward alluded to as MR segments) are utilized. Hadoop, an implementation of the MR model, has been adopted by many companies including the major IT players in the world such as Facebook, eBay, IBM and Yahoo. These implementations are largely done in their respective private clouds. However, recently there are efforts to implement the MR model in public clouds. A major concern of using the MR model in a public cloud is it's in adequate security provision, such as authentication. The MR model was initially intended for use in private networks, so the issue of security was not a design consideration. Since its introduction, lots of efforts have been made to improve the performance of this model making it more efficient rather than making it more secure. Deploying the MR model in an open environment, such as public clouds, without adequate security provisioning would put the clients' jobs and their data at risks. This is because, in such an environment, different jobs submitted by different clients typically share the same set of physical nodes and software resources. The clients have very little control over on which nodes their MR components (assigned to their respective jobs) are executed, and on which DFS nodes the data associated to their jobs are stored.

.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

II. PREVIOUS WORK

One time password or OTP, likewise called vernam-figure or the idealize figure, is a crypto calculation where plaintext is joined with an irregular key. A one-time password is a cryptosystem created by vernam[14]. It's an exceptionally basic framework and is unbreakable if utilized effectively. To utilize an onetime password, you require two duplicates of the "password" which is a square of irregular information measure up to long to the message you wish to encode. One duplicate of the password is kept by every client, and passwords must be traded through a safe channel. The password is utilized by XORing the entire password with all of the first message. Once the message is encoded with the password, the password is demolished and the encoded message is sent. On the beneficiary's side, the encoded message is XORed with the copy duplicate of the password and the plaintext message is created. SecureMR[8] comprises of five security parts, which give an arrangement of down to earth security instruments that not just guarantee MapReduce

benefit honesty and also to forestall replay and Denial of Service (DoS) assaults, yet additionally protect the straightforwardness, pertinence and versatility of MapReduce. We have executed a model of SecureMR based on Hadoop, an open source MapReduce usage. HDFS does not support intra-cloud data encryption yet makes data privacy becomes a key security issue. This paper presents a hybrid encryption method based on HDFS. We adopt symmetric encryption to encrypt and decrypt file blocks at data nodes and use asymmetric encryption scheme to protect the symmetric keys. By this method, we can prevent data node intruders from stealing user data, while ensuring that clients are lightweight. Our security demand is to prevent attackers stealing file data after they intruding into nodes. We insert encryption and decryption modules into nodes. The modules use AES algorithm to encrypt blocks [1] before writing and decrypt before reading. One data block is encrypted and decrypted with a key which is also stored on node. As encryption, decryption and key management modules are deployed at nodes, the modifications to the original protocol between nodes and namenode remains unchanged. The authentication protocol and key exchange protocol is implemented and deployed at clients and nodes.

III. PROBLEM DEFINITION

The MR components can generally be classified into two main categories: master nodes and slave nodes. The Resource Manager and Name Node,are examples of master nodes, and there stare slave nodes. In this version of the MR model implementation, a client submits his job to the Resource Manager. However, The MR components can generally be classified into two main categories: master nodes and slave nodes. The Resource Manager and NameNode, are examples of master nodes, and there stare slave nodes. In this version of the MR model implementation, a client submits his job to the Resource Manager. However, in the classic MR model implementation, a client submits a job to the Job Tracker directly and the Job Tracker then assigns Map and Reduce Tasks to a set of slave nodes. The two sets of MR components, respectively run on two large clusters of nodes are typically referred to as the Processing Framework (PF) cluster and Distributed File System (DFS) cluster. The GMC model is derived to capture the interactions among different MR components in the newer MR model implementation (although what has been captured can also be applied to the classic MR model implementation).More details about the MR components, and their functionalities, of both versions of the MR model implementations (i.e. MR application frameworks) are available. The MR model, owing to its scalability, robustness and simple to use as a parallel and distributed programming framework, is becoming more and more widely used. Hadoop, an implementation of the MR model, has been adopted by many companies including the major IT players in the world such as Facebook, eBay, IBM and Yahoo. These implementations are largely done in their respective private clouds. However, recently there are efforts to implement the MR model in public clouds. A major concern of using the MR model in a public cloud is its in adequate security provision, such as authentication. The MR model was initially intended for use in private networks, so the issue of security was not a design consideration. Since its introduction, lots of efforts have been made to improve the performance of this model making it more efficient rather than making it more secure. Deploying the MR model in an open environment, such as public clouds, without adequate security provisioning would put the clients' jobs and their data at risks. This is because, in such an environment, different jobs submitted by different clients typically share the same set of physical nodes and software resources. In the classic MR implementation, job is submitted by the client to the Job Tracker directly and the Job Tracker then assigns Map and Reduce Tasks to a set of slave nodes. The two sets

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

of MR components, respectively run on two large clusters of nodes are typically referred to as the Processing Framework (PF) cluster and Distributed File System (DFS) cluster . The GMC model, is derived to capture the interactions among different MR components in the newer MR model implementation (although what has been captured can also be applied to the classic MR model implementation).More details about the MR components, and their functionalities, of both versions of the MR model implementations (i.e. MR application frameworks) are available. The MR model, owing to its scalability, robustness and simple to use as a parallel and distributed programming framework, is becoming more and more widely used. Hadoop, an implementation of the MR model, has been adopted by many companies including the major IT players in the world such as Facebook, eBay, IBM and Yahoo. These implementations are largely done in their respective private clouds. However, recently there are efforts to implement the MR model in public clouds. A major concern of using the MR model in a public cloud is its in adequate security provision, such as authentication. The MR model was initially intended for use in private networks, so the issue of security was not a design consideration. Since its introduction, lots of efforts have been made to improve the performance of this model making it more efficient rather than making it more secure. Deploying the MR model in an open environment, such as public clouds, without adequate security provisioning would put the clients' jobs and their data at risks. This is because, in such an environment, different jobs submitted by different clients typically share the same set of physical nodes and software resources.

IV. SYSTEM MODEL

In this section, we briefly explain the model proposed to submit jobs to Hadoop MapReduce framework in a secure manner. This model has been designed so that it provides privacy to each job by authorizing each client and recognizing each job the client submits with a particular job ID. As shown in the Fig 2.1, multiple nodes undergo node authentication which is approved by the Job scheduler. Now clients are allowed to submit their jobs after they are authenticated with the help of client authenticator. Once they are authenticated, they can submit their jobs to the job processor. These jobs are scheduled with the help job scheduler. This is done is a modular fashion i.e, when the jobs submitted by the different clients are more than the number of working nodes, the jobs are allotted to the nodes using modularity. The nodes will send the output data to the Data Store which in turn encrypts the output data and also decrypts the data only when the authenticated client asks to view the result of the job using the Data Handle.

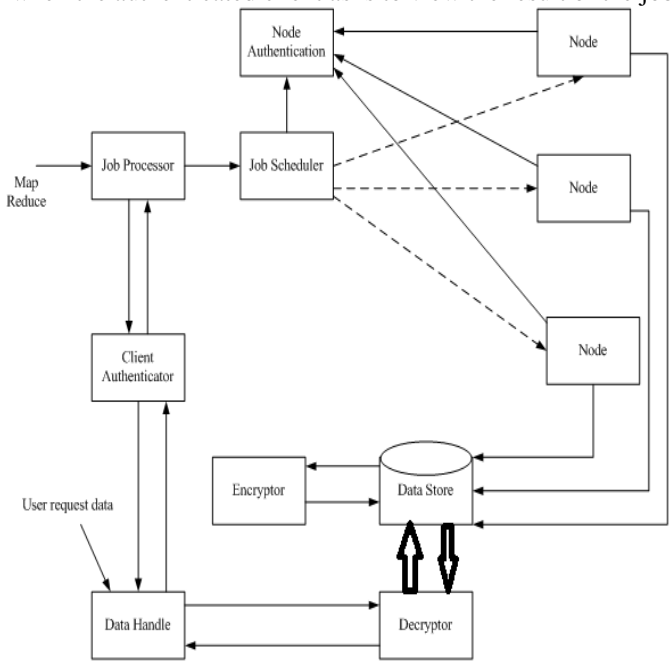


Fig. 2.1 Overall system architecture of a secure model for client authentication in MR framework

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

V. PROPOSED METHODOLOGY

In the proposed Methodology, for implementation purpose, Java is chosen as the programming language. The entire implementation is programmed in java using the Hadoop MapReduce framework and the HDFS architecture is used to encrypt and decrypt the data generated by the jobs submitted by the clients. With the implementation of AES algorithm, the data generated will be encrypted and decrypted. The user interface is generated using the Java swing functionalities. Three separate swing portals have been created for the client and node processing. MAC code is generated for each node through which the client can allot jobs to a particular node and have a secure communication. When the number of jobs submitted by the clients is more than the nodes for created, the application just uses the function of modularity to allot more than one job to different nodes.

VI. EXPERIMENTAL RESULTS

The proposed model uses an efficient way to allot the jobs from different clients to different nodes. It uses modular arithmetic to allot jobs to nodes (as mentioned in the section II). This consumes a time complexity of O(1). Job scheduling does not take much time as it does not involve any security tasks. Security is provided to their output data at the data storage end. This model also uses the one time password methodology which also optimizes the time efficiency

VII. FUTURE SCOPES

The model which we have proposed above has certain limitations. This model enables accepting of jobs from a single client at a time. So, we can enhance it's functionality by extending the model's feature to accept multiple jobs from multiple clients simultaneously. This work can further increase the parallelization of job processing.

VIII. CONCLUSION

The proposed solution for the MR Job processing is being implemented in fashion as mentioned in the system architecture diagram, and we have proved that this architecture is better model for providing security and privacy of each client and their jobs submitted to each data node.

REFERENCES

[1] J. Dyer and N. Zhang, ''Security issues relating to inadequate authentication in MapReduce applications,'' in Processor Int. Configuration High Performance Comput. Simulation (HPCS), Jul. 2013,pp.281–288.
[2] T. White, ''How the MapReduce works,'' in Hadoop: The Definitive Guide, 3rd ed. Tokyo, Japan: O'Reilly Inc., 2012.
[3] I. Lahmer and N. Zhang, ''MapReduce: MR model abstraction for future security study,'' in Proc. 7th Int. Conf. Secur. Inf. Netw., 2014,pp.392–398.
[4] C. Lam, ''Introducing Hadoop, and managing hadoop,'' in Hadoop in Action. Greenwich, U.K.: Manning Publications Co, 2010.
[5] P. Zikopoulos, C. Eaton, D. Deroos, T. Deutsch, and G. Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: McGraw-Hill, 2012.
[6] J. Dean and S. Ghemawat, ''MapReduce: Simplified data processing on large clusters,'' Commun. ACM, vol. 51, no. 1, pp. 107–113,2008.
[7] J. Xiao and Z. Xiao, ''High-integrity MapReduce computation in cloud with speculative execution,'' in Theoretical and Mathematical Foundations of Computer Science. Heidelberg, Germany: Springer-Verlag,2011,pp.397–404.
[8] B. Lakhe, ''Introducing Hadoop and its security,'' in Practical Hadoop Security. New York, NY, USA: Apress, 2014.
[9] I. Lahmer and N. Zhang, ''MapReduce: A security analysis and authentication requirement specification,'' in Proc. 2nd Int. Conf. Comput. Inf. Syst. (ICCIS), World Congr. Comput. Appl. Inf. Syst., 2015,pp.65–71.
[10] D. A. B. Fernandes, L. F. B. Soares, J. V. Gomes, M. M. Freire, and P. R. M. InÆcio ''Security issues in cloud environments: A survey,'' Int. J. Inf. Secur., vol. 13, no. 2, pp. 113–170, Apr. 2014.
[11] J. M. Kizza, ''Cloud computing and related security issues,'' in Guide to Computer Network Security. London, U.K.:Springer-Verlag,2013,pp.465–489.
[12] A. Kumar, S. Jakhar, and S. Makkar, ''Comparative analysis between DES and RSA algorithms,'' Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 2, no. 7, pp. 386–391, Jul. 2012.
[13] O. O'Malley, K. Zhang, S. Radia, R. Marti, and C. Harrell, ''Hadoop security design,'' Yahoo, Inc., Sunnyvale, CA, USA, Tech.Rep.,2009.
[14] N. Somu, A. Gangaa, and V. S. S. Sriram, ''Authentication service in Hadoop using one time pad,'' Indian J. Sci. Technol., vol. 7, pp. 56–62, Apr. 2014.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

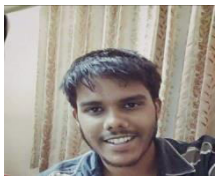
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

[15] S. Rubika, G. S. Sadasivam, and K. A. Kumari, ''A novel authentication service for Hadoop in cloud environment,'' in Proc. IEEE Int. Conf. Cloud Computing Emering. Markets (CCEM), Oct. 2012, pp. 1–6

BIOGRAPHY



Chetan Kashyap, final year B. E., Department of Computer Science & Engineering, P.E.S. College of Engineering, Mandya, Karnataka

Puneeth H. V, final year B. E., Department of Computer Science & Engineering, P.E.S. College of Engineering, Mandya, Karnataka

Sukruth R, final year B. E., Department of Computer Science & Engineering, P.E.S. College of Engineering, Mandya, Karnataka

Prajwal M. H, final year B. E., Department of Computer Science & Engineering, P.E.S. College of Engineering, Mandya, Karnataka

Raghavendra Babu .T.M (Guide), Assistant Professor, Department of Computer Science & Engineering, P.E.S. College of Engineering, Mandya, Karnataka