Designing Authentication for Hadoop Cluster using DNA Algorithm.

Balaraju.J., PVRD. Prasada Rao

Abstract: Big Data (BD) generation are exponentially increased and it is necessarily required for modern society. Hadoop Clusters (HC) provides the facilities like Processing, storage and doesn't have built-in security. But this feature is important as it increases the analysis speed, storage process. HC facilitates storage, processing of data, on the other hand processing of streaming data handled by the Apache Spark. However data storage, processing power, cluster management and data security in HC is not reached up to the mark with increased data. In such situations, HC are scaled out from small scale IT organization and it depends on public cloud centers with lack of data security, communication, computation and operational cost. On the other hand data security in HC is major issue and it uses a separat security mechanisms. This paper proposes New Algorithm Built in Authentication Based on Access (BABA) as a security instance integrated as Hadoop instance for securing data in HC from attackers along with metadata security for avoiding crashes Hadoop. This mechanism provides a secured HC without using other security configurations which will reduce operational cost, computational power, increases data security and providing a better solution for HC.

Keywords: Big Data, Authentication, Hadoop Cluster, Access System, Security.

I. INTRODUCTION

Big Data [1] generation is high and handling is intuition task by using traditional technologies. Since, BD plays a crucial role in modern society, it requires a secured distributed environment by providing scalable storage, analysis of static and streaming data. Advanced technology like Hadoop [2] provides a scalable and distributed storage, parallel processing framework for data analysis to BD. Hadoop Clusters [3] provides huge data storage and analysis of huge amount of unstructured data in distributed passion with parallel processing environment by supporting commodity hardware. HCs are highly scalable, by providing boosting of data analysis application. HCs are flexible for adding of processing power when data generation is increased by adding extra node in to the cluster. HCs are high resistant to failure of data, each piece of data is copied in multiple nodes in different racks. All JVM instances in Hadoop Single Node cluster are runs in only one machine having less processing power and storage capacity by replicating one copy of each data. JVM in Hadoop Multi Node Cluster are runs in different nodes like master nodes, slave and client nodes with high processing power, huge storage capacity with user defined replicas are supports and its default replica is 3.

II. MAJOR SERVICES IN HADOOP CLUSTER.

A. Master Node.

MasterNodes [4] in distributed HCs provides management services like storage of large data in distributed manner and parallel processing of huge data by using HDFS [5] and MapReduce[6] services. Master Nodes like NameNode [7] is designed for managing storage of HDFS by storing users metadata, SecondaryNameNode [8] is also checkpoint Node or backup Node for security of metadata in the absence of NameNode. JobTracker [9] or Resource manager or YARN is available later versions of hadoop2.0 which is act as resource manager for parallel processing of data to speeding up the data analysis.

B. Slave Node [10]

In HC slave nodes are used for storage of data and facilitating data analysis by running every tasks with in their processor. Each slave node manages individual map or reduce task as a TaskTracker [11] in hadoop1.0 and it is replaced as a YARN services in Hadoop 2.0 later version. DataNode [12] is provides a service to stores huge data or running MapReduce operations. The DataNodes are communicating with NameNode in regular intervals for updating metadata and also the TaskTracker communicating with JobTracker regularly.

C. Client Nodes

Client nodes are providing gateway to HC and outer networks are used for running cluster administrative tools as well as client application. Edge node / client node / Gatway Node is not part of the hadoop cluster and it does not runs any hadoop services. Edge node uses hadoop binaries and hadoop cluster configured files to submitting jobs on HC. Hadoop Multimode cluster is having facility for configuring Edge node by installing client tools like HIVE, SQOOP, PIG, and OOZIE.

III. HADOOP CLUSTERS

A. Single Node Cluster (SNC).

Hadoop Single Node Cluster (SNC) [13] is non distributed or standalone mode with single java instance runs for all daemons in one node. The Hadoop Multi Node Cluster (MNC) [14] can have more than one DataNodes. In a SNC, all the daemons [15] like DataNode, NameNode, TaskTracker and JobTracker are configured with one dedicated server with limited process power. Data storage and processing is limited in SNC without security futures, a separate Security mechanism is required for securing the data, data going to be loss when server is crashed. SNC maintained only one copy of replica by default with in the server.

Manuscript published on 30 September 2019
* Correspondence Author

Mr.J.Balaraju*, Assistant Professor in the Computer Science & Engineering Department at RGM College of Engineering & Technology
Dr PVRD Prasada Rao, Professor in the Computer Science & Engineering Department at KL University.

Retrieval Number: C5895098319/2019©BEIESP
DOI:10.35940/ijrte.C5895.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A single JVM instance runs for data storage and processing in standalone node on top of the OS.

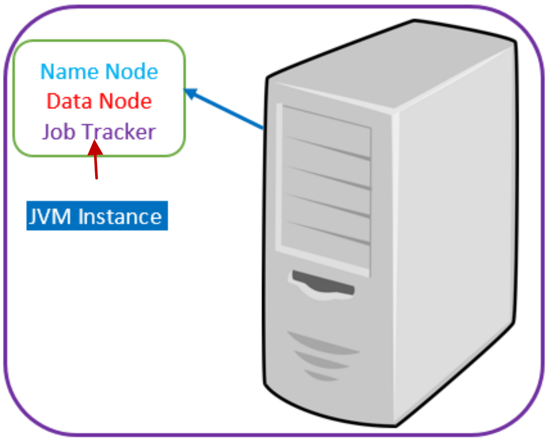


Figure 1: Typical Single Node HC.

B. Multi Node Cluster (MNC).

Hadoop MNC provides a distributed, parallel processing environment and it contains more than one DataNodes with huge data storage capacity and processing power. All the daemons are completely runs on different nodes as master and slave nodes with default replication factor is 3 which provides high security future than SNC. A MNC setup is look like master slave architecture among that one machine acts as a master that runs the NameNode, other one as JobTrackers / YARN daemons, whereas other nodes are acts as slave nodes for data storage and processing as DataNodes. Hadoop MNC is facility for supporting commodity hardware that run the TaskTracker and DataNode daemons whereas alternative services are run on High configure servers. Hadoop based MNC is suitable for storage of large data, speed up data analysis and its uses a third party security as separate configurations which is the major disadvantage of Hadoop MNC.

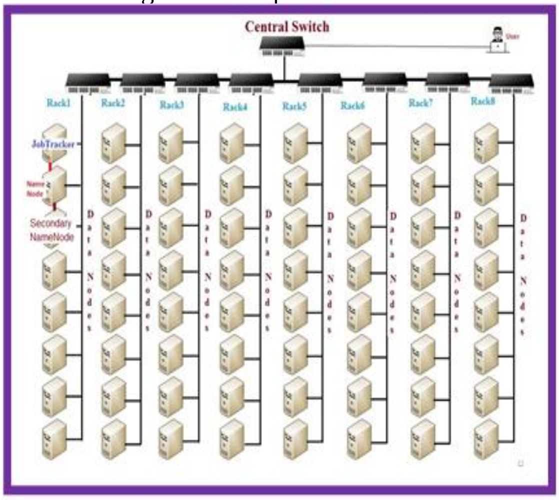


Figure 2: Typical MultiNode HC.

IV. PROBLEM DEFINITION.

Data storage and utilization are required for every need and it is necessary to maintain dedicated data center in every organization. Building the Hadoop based MNC is very costlier and it requires high establishment cost, operational maintenance and MNC is not necessary for small organization. Most of the organizations are depending on public cloud data center for their regular storage and processing purpose. This can be a lack of computational, communicational and data security. Based on the above challenges it is recommended to establish a Hadoop based Secured Cluster with large data storage and increases the speed of data analysis.Though Hadoop framework is popular for Data storage and analysis, it has the following security problems.

- DataNode Does not have control over their Data Blocks.
- Lack of Metadata Security. If NameNode and Secondary NameNode crash or not Available, scope for HC crash.
- Master Services Does Not enforce any user for Authentication.

On the other hand,the problem with Hadoop based SNC is data security as a vital task, because all instances (JVM) runs in single Node and all users having equal access permissions with all instances. The instances which are NameNode storing metadata, JobTracker handling user’s jobs and DataNode stores the original data in their blocks. Any user can access data directly from DataNode without authentication. In any HC NameNode is not available user cannot access their data without metadata ultimately finally there is a scope for Hadoop crash. Hadoop does not have its own security mechanism and it depends on other instances like Kerberos [16] for authentication, Apache Knox [17] is single point authentication, Apache Ranger [18] and Ranger KMS [19] is used for Data encryption purposes. Based on above challenges, Hadoop Clusters required its own security protocol for increasing utilization Hadoop based SNC or MNC which increases in data security performance and decreases the operational, communicational cost for every organization.

V. EXISTING SECURITY MECHANISM.

Literature Survey:

Ahlam Kourid , Salim Chikhi et al [20] are discussed recent advances in Big Data for Security and Privacy in different levels like network, data application and authentication level based on 5V characteristics of big data domain. Authors also discussed a relative study of existing security mechanism using hadoop and given 7 future direction for big data security. One among that, by adding extra layer in HC or security instance in hadoop framework is provides a solution for big data security.

Balaraju.J, Dr.P.V.R.D. Prasada Rao et al [21] are implemented a secure authentication mechanism by adding Secure-DNA node in HDFS along with metadata security. They are also discussed different security techniques and proposed a good solution for big data security for Hadoop based cloud center by integrating their implementation by using data hiding technique like DNA.

VI. CONTRIBUTION OF MY WORK – NEW APPROACH USING DNA ALGORITHM.

Deoxyribonucleic acids (DNA)[22] is best data hiding technique compare to other data hiding techniques. DNA gives indirect security mechanism and its sequence difficult to understand hackers. DNA haves four types of proteins namely Adenine (A) Cytosine(C), Guanine (G) and Thymine (T) and each proteins represents two digit binary numbers in the following manner.

Proteins Name	Binary Digits
Adenine (A)	00
Cytosine(C)	01
Guanine (G)	10
Thymine(T)	11

DNA cryptography gives more interest to scientist for developing new DNA based security algorithms because of its complex structure and no direct connection between data and DNA sequence. Most of the existing DNA based security algorithms are developed for data and image encryption and decryption but in this paper we used DNA sequence for hiding of sensitive data for both authentication as well as providing access permission to authorized user. Designed and developed the following algorithms based on the above techniques among the two are generating user mail id to Unique key and reverse for user authentication and two are Nodes unique property like MAC Address which is in Hexa Decimal form by generating Unique key and reverses. This algorithm also useful for finding cluster size, status of the node, adding and deletion of nodes from the cluster. Final algorithm is main for providing Access Data to authorized users to access data from DataNode. Datanode have full control over the existing data blocks which is major disadvantage in Hadoop.

Algorithm for Data Hiding

MailidtoUniqueKey ()

```
{
    $M\_Size \leftarrow strlen(mailid)$
    for i in 1 to M_Size
        {
    $ASC_i \leftarrow mailid_i$.
    $BF_i \leftarrow Convert\ ASC_i$.
    $DNAF_i \leftarrow BF_i$.
    $AV_i \leftarrow DEC_i$    // A=0,C=1,G=2,T=3.
    $DEC_i \leftarrow Sum\ of\ Quadruple_i$.
    $U\_Key \leftarrow DEC$
        }
}
```

UniqueKeytoMailid ()

```
{
    $Ukey\_Size \leftarrow Count(U\_Key)$.
    for i in 1 to UKey_Size
        {
    $DEC_i \leftarrow Sum\ of\ Quadruple_i$.
    $AV_i \leftarrow DEC_i$    // A=0,C=1,G=2,T=3.
    $DNAFi \leftarrow AV_i$
    $BF_i. \leftarrow DNAFi$
    $ASCi \leftarrow BF_i$
    $mailid \leftarrow ASCi$.
        }
}
```

MACAddresstoUniqueKey ()

```
{
Begin
    $Cluser\_Size \leftarrow strlen(MAC\_Addr)$
    for i in 1 to Cluster_Size
        {
    $BF_i \leftarrow Convert\ MAC\_Addr_i$.
    $DNAF_i \leftarrow BF_i$.
    $AV_i \leftarrow DEC_i$    // A=0,C=1,G=2,T=3.
    $DEC_i \leftarrow Sum\ of\ Quadruple_i$.
    $U\_Key \leftarrow DEC$
        }
}
```

UniqueKeytoMACAddress ()

```
{
    $Ukey\_Size \leftarrow strlen(U\_Key)$.
    for i in 1 to UKey_Size
        {
    $DEC_i \leftarrow Sum\ of\ Quadruple_i$.
    $AV_i \leftarrow DEC_i$    // A=0,C=1,G=2,T=3.
    $DNAFi \leftarrow AV_i$
    $BF_i. \leftarrow DNAFi$
    $MAC\_Addr_i \leftarrow BF_i$
        }
}
```

The proposed system Built Authentication Based on Access (BABA) is integrated as a JVM instance along with existing bundle of instances as agent system for securing the data in Hadoop based SNC. BABA act as security interface between users and data in SNC in regular intervals by watching user activities. Every user or organization must register with BABA by sending credentials and selecting a unique property like mail id for the first time. BABA generates a unique key for user for authentication and this unique key also sending to all instances for verifying the regular activities in future. All the user's unique keys are hided by using hiding technologies like DNA Cryptography and it's not visible by other user. BABA also maintain two columns table by storing the unique key in one column, users metadata in second column by collecting from NameNode in regular interval and sends to users mail for future accessing their data from DataNode in the absence of NameNode which lead crashes of Hadoop SNC.

VII. SECURITY ENHANCEMENT HADOOP SNC WITH BABA.

In any cluster the initial data is kept in memory blocks of DataNode and it doesn't have management over them. In SNC the DataNode is a JVM instance by collecting free HDD space with 128 MB to 512MB block size which is 64 MB earlier version.

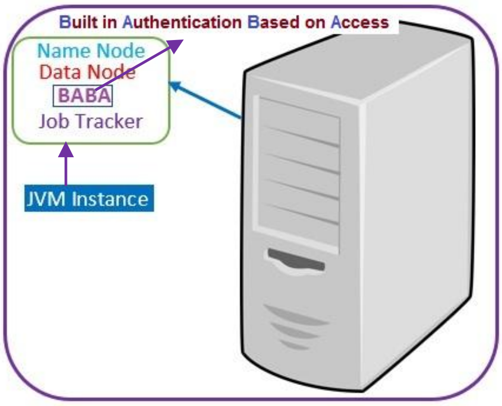


Figure 3.: Proposed SNC with BABA.

BABA provides a partial accessing privileges for running DataNode instance, any user wants to access data blocks from

DataNode it's mandatory user must have Unique key otherwise it rejects the user. Once user is authorized they can have privileges to access data from blocks for limited time. User access time is exceeded more than assigned time, they want specify approximate time for working with data blocks initially. The authorized user have full control over its own data and partial control on others data for securing the data. The proposed one can be provide a control over data block for DataNode ultimately unauthorized users cannot access data directly from DataNode.

VIII. SECURITY CONTRIBUTION HADOOP MNC WITH BABA.

Hadoop MNC supports more number of nodes approximately 10300 by dividing master and slave services and many users are work with in the cluster. Any user can enter in to the cluster with equal privileges with the help of master services and this services does not enforce user authentication.

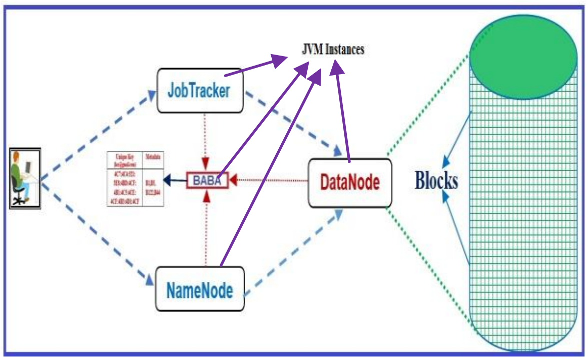
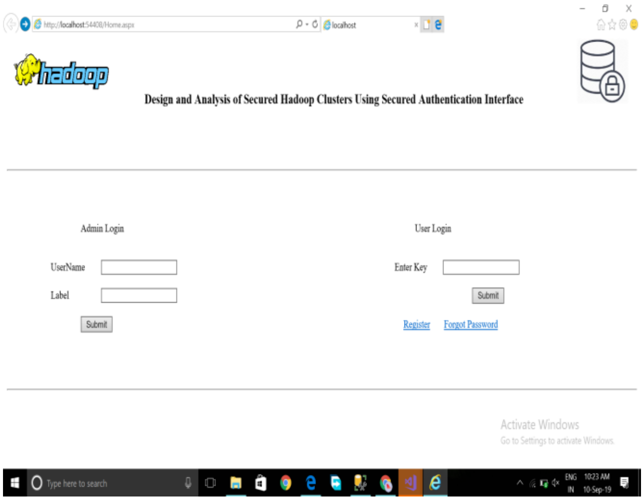


Figure 4.: Hadoop MNC with BABA

In MNC Master Services runs in different high end servers and remaining nodes in the clusters are DataNodes. BABA act as a Data guardrail to all the DataNodes of HC for verifying users authorization at the time of data access and other activities performs on DataNodes. BABA provides completes privileges to DataNodes over their DataBlocks which provides better security of BD.

IX. RESULTS AND ANALYSIS.


http://localhost:54408/Home.aspx

Design and Analysis of Secured Hadoop Clusters Using Secured Authentication Interface

Admin Login

UserName

Label

Submit

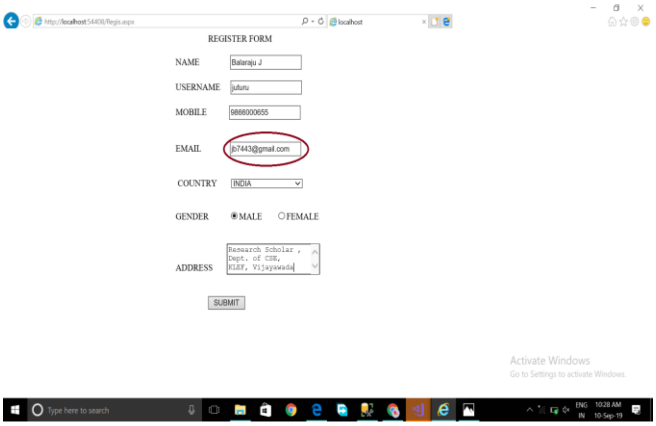
User Login

Enter Key

Submit

Register Forgot Password

Fig 5. Secured Authentication Interface.


http://localhost:54408/Regis.aspx

REGISTER FORM

NAME Balaraju J

USERNAME juturu

MOBILE 9866000655

EMAIL jb7443@gmail.com

COUNTRY INDIA

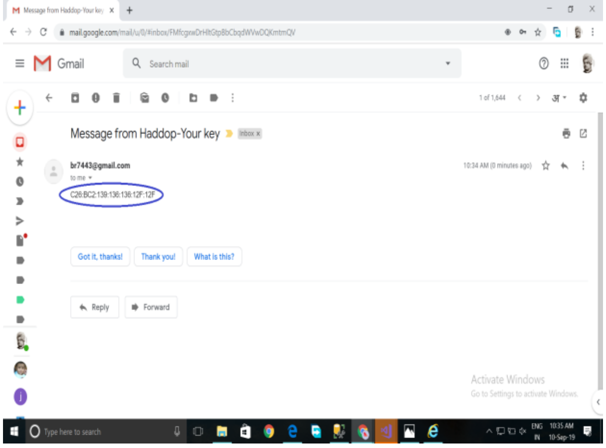
GENDER ● MALE ○ FEMALE

ADDRESS Research Scholar ,
Dept. of CSE,
KLEF, Vijayawada

SUBMIT

Fig 6. User Registration Page.

In this phase user mail id is mandatory for generating permanent static key which is using DNA algorithm. Once submitting the registartion form automatically key will sending to users mail id.


mail.google.com/mail/u/0/#inbox/FMfcgxwDrHltGtpBbCbqdWVwDQKmtmQV

Message from Haddop-Your key Inbox

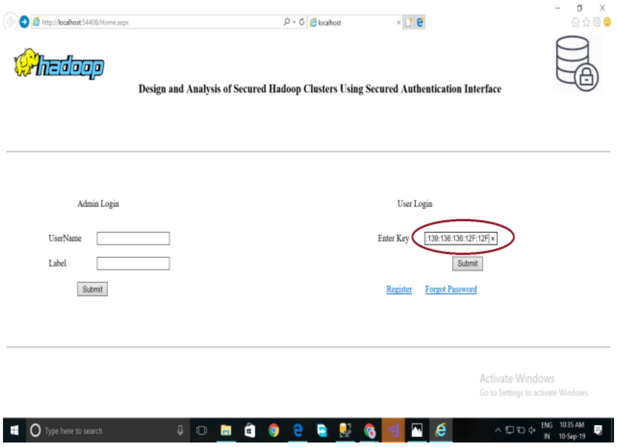
br7443@gmail.com 10:34 AM (0 minutes ago)
to me

C26:BC2:139:136:136:12F:12F

Got it, thanks! Thank you! What is this?

Reply Forward

Fig 7. Users inbox with generated key.


http://localhost:54408/Home.aspx

Design and Analysis of Secured Hadoop Clusters Using Secured Authentication Interface

Admin Login

UserName

Label

Submit

User Login

Enter Key :139:136:136:12F:12F

Submit

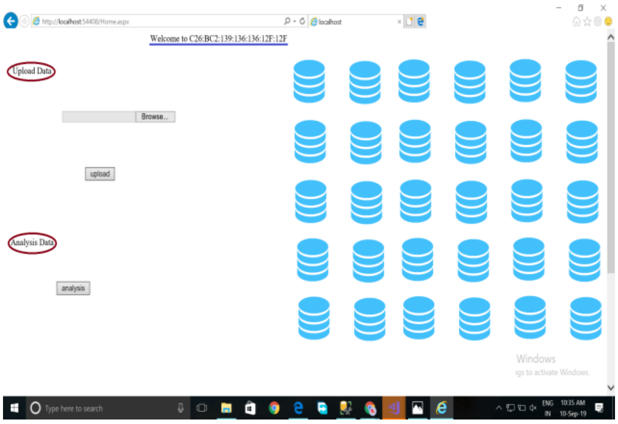
Register Forgot Password

Fig 8: Login Page with Generated Key.

Only valid key is used for login in to the Hadoop cluster for first time and seocond time onwords dynamic added to the static key which will be generationg once exit from the cluster.

Retrieval Number: C5895098319/2019©BEIESP
DOI:10.35940/ijrte.C5895.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



http://localhost:54408/Home.aspx

Welcome to C26:BC2:139:136:136:12F:12F

Upload Data

Browse...

upload

Analysis Data

analysis

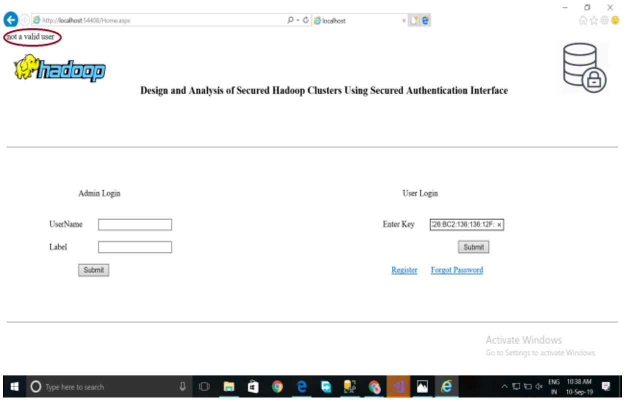
Windows
ngs to activate Windows.

Type here to search

10:35 AM
10-Sep-19

Fig 9.Hadoop cluster with existing Nodes.

Once user login is successfully login in to HC each and every user events are saved by interface. Users have two options Upload data option is for storing the data and process data is analysis of data. At the user logout time this interface generating second part of the dynamic key and user must be login with two keys next time onwards.



http://localhost:54408/Home.aspx

not a valid user

Design and Analysis of Secured Hadoop Clusters Using Secured Authentication Interface

Admin Login

UserName

Label

Submit

User Login

Enter Key :26:BC2:136:136:12F:

Submit

Register Forgot Password

Activate Windows
Go to Settings to activate Windows.

Type here to search

10:38 AM
10-Sep-19

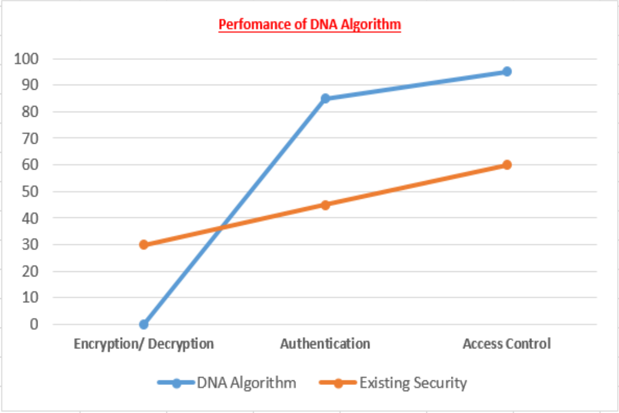
Fig 10. Login Failed

X. SECURITY PERFORMANCE.

This algorithm likely provides 24X7 security for Hadoop Cluster which is very useful for small organizations to maintain their data center without depending on public or private data center which are for away to the organization. This can increase the data security, communication operational, and reduce maintenance problems.

Performance of DNA algorithm with existing technologies.

Security Attribute	Authentication.	Access Control.	Data Encry / Decr
Kerberos	√		
Apache Knox(SP)	√		
Apache Ranger		√	
Ranger KMS			√
DNA Algorithm (Proposed)	√	√	



In Addition to this, there is a scope in MNC for finding the status of node, adding extra node and deletion of node from cluster is easy by using this algorithm.

XI. CONCLUSION AND FUTURE WORK.

The proposed algorithm mainly concentrates on providing control over data blocks for DataNode which actually stores the original data. This algorithm also provides the security of metadata by sending that to authorized users in regular intervals. Users can access their data in absence of NameNode and Secondary NameNode with the permission of DataNode, ultimately this can reduces the crashing of Hadoop. A dedicated JVM secure instance is better solution for HC without using other security mechanisms. By using this, there is a scope for configuring SNC in organization by reducing operational, computational cost and increasing data security and provides better security for MNC's. The enhancement of this work is to reducing the computational burdens of the proposed algorithm.

REFERENCES:

1. Ishwarappa, & Anuradha, J. "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology". Procedia Computer Science, 48, 319–324. doi:10.1016/ j. procs .2015 .04.188.
2. Nandimath, J., Banerjee, E., Patil, A., Kakade, P., Vaidya, S., & Chaturvedi, D. "Big data analysis using Apache Hadoop". IEEE 14th International Conference on Information Reuse & Integration (IRI). doi:10.1109/iri.2013.6642536.
3. Jiong Xie, Shu Yin, Xiaojun Ruan, Zhiyang Ding, Yun Tian, Majors, J., Xiao Qin. "Improving MapReduce performance through data placement in heterogeneous HCs". IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW). doi:10.1109/ ipdpsw.2010 .5470880.
4. Ren, Z., Xu, X., Wan, J., Shi, W., & Zhou, M. (2012). Workload characterization on a production HC: A case study on Taobao. IEEE International Symposium on Workload Characterization (IISWC). doi: 10.1109 /iiswc. 012. 6402895.
5. Shvachko, K., Kuang, H., Radia, S., & Chansler, R. "The Hadoop Distributed File System". IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). doi:10.1109/ msst.2010. 5496972.
6. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Commun. ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008.
7. Mackey, G., Sehrish, S., & Wang, J. "Improving metadata management for small files in HDFS".IEEE International Conference on Cluster Computing and Workshops. doi:10.1109/ clustr.2009.5289133.

8. Won, H., Nguyen, M. C., Gil, M.-S., Moon, Y.-S., & Whang, K.-Y. "Moving metadata from ad hoc files to database tables for robust, highly available, and scalable HDFS". The Journal of Supercomputing, 73(6), 2657–2681. Doi: 10.1007/ s11227-016-1949-7.
9. Kim, Y.-P., Hong, C.-H., & Yoo, C. "Performance impact of JobTracker failure in Hadoop". International Journal of Communication Systems, 28(7), 1265–1281. Doi: 10.1002 / dac. 2759.
10. Yao, Y., Wang, J., Sheng, B., Tan, C. C., & Mi, N. "Self-Adjusting Slot Configurations for Homogeneous and Heterogeneous HCs". IEEE Transactions on Cloud Computing, 5(2), 344–357. doi:10.1109/ tcc.2015. 2415802.
11. Cheng, D., Rao, J., Guo, Y., Jiang, C., & Zhou, X. "Improving Performance of Heterogeneous MapReduce Clusters with Adaptive Task Tuning" . IEEE Transactions on Parallel and Distributed Systems, 28(3), 774–786. doi:10.1109/ tpds.2016. 2594765.
12. Park, D., Kang, K., Hong, J., & Cho, Y. "An efficient Hadoop data replication method design for heterogeneous clusters". Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16. doi:10.1145/2851613. 2851945.
13. Pradhananga, Y., Karande, S., & Karande, C. "High performance analytics of bigdata with dynamic and optimized HC". International Conference on Advanced Communication Control and Computing Technologies (ICACCCT). doi:10.1109/ icaccct. 2016.7831733.
14. Singh, R., & Kaur, P. J. "Analyzing performance of Apache Tez and MapReduce with hadoop multinode cluster on Amazon cloud". Journal of Big Data, 3(1). doi:10.1186/s40537-016-0051-6.
15. Ghazi, M. R., & Gangodkar, D. "Hadoop, MapReduce and HDFS: A Developers Perspective" . Procedia Computer Science, 48, 45–50. doi:10.1016 /j.procs.2015.04.108.
16. Pedro Camacho, Bruno Cabral, Jorge Bernardino. "Insider Attacks in a Non-secure Hadoop Environment" WorldCIST 2017: Recent Advances in Information Systems and Technologies pp 528-537.
17. Singh, U., Kumar Solanki, N., Kumar Varma, M., & Sevak, T. "A review on big data protection of Hadoop". 2nd International Conference on Communication and Electronics Systems (ICCES). doi:10.1109/cesys.2017 .8321221.
18. Revathy, P., & Mukesh, R. "Analysis of big data security practices". 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). doi:10.1109/ icatcct.2017.8389145.
19. Gupta, S., & Giri, V. "Data Security in Data Lakes". Practical Enterprise Data Lake Insights, 225–259. doi:10.1007/978-1-4842-3522-5_6.
20. Ahlam kourid and Salim Chikhi " A comparative Study of Recent Advances in Big Data for Security" P.No:249-258 , Networking Communication and Data Knowledge Engineering, Springer Nature Singapore 2018.
21. Balaraju.J, Dr.P.V.R.D. Prasada Rao, "Enhanced Security For Hadoop Distributed File System By Using DNA Cryptography" , in International Journal of Pure and Applied Mathematics Volume 120 No. 6 2018, 8127-8142 ISSN: 1314-3395.
22. Wang, B., Xie, Y., Zhou, S., Zhou, C., & Zheng, X. "Reversible Data Hiding Based on DNA Computing. Computational Intelligence and Neuroscience", 2017, 1–9. Doi: 10.1155/ 2017/ 7276084.

AUTHORS PROFILE

Mr.J.Balaraju received his M.Tech. in Computer Science from JNTUA, Ananthapuramu in 2012. He is a Assistant Professor in the Computer Science & Engineering Department at RGM College of Engineering & Technology (Autonomous), Nandyal and Research Scholar in Koneru Lakshmaiah Educational Foundation (Deemed To be University), Vijayawada . His research areas include big data analytics,data mining,,IoT and Sensor network.

Dr PVRD Prasada Rao received his Ph.D.in Computer Science from Acharya Nagarjuna University in 2014 and M.Tech in CSE from Andhra University in 1998. He is a Professor in the Computer Science & Engineering Department at KL University.

His research areas include data mining, bioinformatics,IoT, Sensor network and big data analytics. Dr PVRD Prasada Rao has published 70+ research papers in the leading international journals and conference proceedings. In addition he is a Associate Dean(P&P)/Reviewer/Member of several international conferences/workshops.