A Novel Node Management in Hadoop Cluster by using DNA

Balaraju.J[1†]. and Dr.PVRD.Prasada Rao[2††].

Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

Abstract.
The distributed system is playing a vital role in storing and processing big data and data generation is speedily increasing from various sources every second. Hadoop has a scalable, and efficient distributed system supporting commodity hardware by combining different networks in the topographical locality. Node support in the Hadoop cluster is rapidly increasing in different versions which are facing difficulty to manage clusters. Hadoop does not provide Node management, adding and deletion node futures. Node identification in a cluster completely depends on DHCP servers which managing IP addresses, hostname based on the physical address (MAC) address of each Node. There is a scope to the hacker to theft the data using IP or Hostname and creating a disturbance in a distributed system by adding a malicious node, assigning duplicate IP. This paper proposing novel node management for the distributed system using DNA hiding and generating a unique key using a unique physical address (MAC) of each node and hostname. The proposed mechanism is providing better node management for the Hadoop cluster providing adding and deletion node mechanism by using limited computations and providing better node security from hackers. The main target of this paper is to propose an algorithm to implement Node information hiding in DNA sequences to increase and provide security to the node from hackers.

Keywords:
Distributed System, Hostname, Hadoop, MAC, Node, DNA sequences.

1. Introduction.

Distributed Computing (DC) [1] furnishes a practical edge work with productive execution of an answer on various PCs associated with a system. For conveyed Computing DC, huge undertakings are partitioned into littler issues which would then be able to be executed on various PCs simultaneously free of one another. The assignment must be separated into free issues to limit the PC's correspondence. In any case, dispersed figuring won't be powerful. In the course of recent years, the intermixing of software engineering and the multifaceted nature of science has lead to the prosperous field of bioinformatics. Advances in atomic science and innovation for huge bits of genomes in different species. Today PCs have made clinical examination more effective and exact, by utilizing equal and conveyed PCs and complex organic displaying. Bioinformatics [2] is one of the more up to date territories and has made us fully aware of a totally different universe of science. The combination of PCs and science has helped researchers study species, particularly people. With the guide of the PCs, we have taken in a lot about hereditary qualities, however, there still stand numerous unanswered inquiries that are being investigated today. Deoxyribonucleic

Acid (DNA) [3] arrangement investigation can be a protracted cycle going from a few hours to numerous days. This paper manufactures a dispersed framework that gives the answer for some bioinformatics related applications. The general objective of this paper is to assemble a Distributed Bioinformatics Computing System for hereditary succession examination of DNA. This framework is equipped for looking and distinguishing quality examples in a given DNA grouping. To process, we put away a huge no. of DNA succession utilizing Different lengths of DNA arrangements were utilized for the sequential and nonconsecutive example search to think about the framework's reaction time acquired utilizing single and numerous PCs. Furthermore, various lengths of DNA arrangements were likewise utilized for the example ID to think about its reaction time watched utilizing a solitary PC and numerous PCs. A few diverse disseminated usages of search calculations have been accounted for in the writing. It tends to be seen that the majority of the current methodologies require superior equal processors and are not executed on inexactly coupled conveyed organize. Besides, the majority of them require specific programming language for their execution on these equal processors. The particular target of the proposed disseminated calculation for investigation of DNA successions are 1. Build up a viable disseminated DNA arrangement examination calculations for design coordinating of DNA Gene grouping and sub-successions recognizable proof. 2. Execute them on an approximately coupled appropriated system, for example, customary neighborhood and wide

Manuscript received September 5, 2023
Manuscript revised September 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.9.17

territory organize utilizing standard programming language.

The principal focus of this paper is to propose a calculation to actualize information covering up in DNA groupings to expand the multifaceted nature and making disarray to the programmers. By using some fascinating highlights of DNA arrangements, the usage of data stowing away is applied. The calculation which has been proposed here depends on parallel coding and the corresponding pair rules.

2. Hadoop Cluster.

Hadoop [4] is an Apache open source system written in java that permits dispersed preparing of huge datasets across bunches of PCs utilizing straightforward programming models. The Hadoop structure application works in a situation that gives dispersed capacity and calculation across groups of PCs. Hadoop is intended to scale up from single worker to a great many machines, each offering neighborhood calculation and capacity. In small Hadoop Cluster (HC) [5] have a solitary ace Node Server and various customer hubs. The ace hub comprises of a Job Tracker, Task Tracker, NameNode, and DataNode. A slave or specialist hub goes about as both a DataNode and TaskTracker, however it is conceivable to have information just and figure just laborer hubs. In a bigger HC, HDFS hubs are overseen through a committed NameNode worker to have the document framework record, and an optional NameNode that can produce depictions of the namenode's memory structures, accordingly forestalling document framework debasement and loss of information The size of HC,s are quickly expanded from 2005, the inventers is restricted to just 20 to 40 nodes in a groups. At that point they understood two issues, they are not accomplish its potential until it ran dependably on the bigger groups. In the second stage Yahoo effectively tried Hadoop on a 1000 hub bunch and begin utilizing it later yippee and Apache Software Foundation effectively tried a 4000 hub group with Hadoop. HC bunch expanded 4000 to 10000+ in various deliveries.

3. Roles of Dynamic Host Configuration Protocol (DHCP).

DHCP [6] is a convention that gives snappy, programmed, and focal administration for the dispersion of IP addresses [7] inside a system. DHCP is additionally used to arrange the subnet cover, default entryway, and DNS worker data on the device. Media Access Control (MAC) Address is a novel identifier of the Network Interface Controller (NIC). A system hub can have different NIC yet each with novel MAC. A system chairman saves a scope of IP addresses for DHCP, and each DHCP customer on the LAN is arranged to demand an IP address from the DHCP worker during system introduction. The solicitation and-award measure utilizes a rent idea with a controllable timeframe, permitting the DHCP worker to recover and afterward reallocate IP tends to that are not recharged. The DHCP worker forever appoints an IP address to a mentioning customer from the range characterized by the executive. This resembles dynamic allotment, however the DHCP worker keeps a table of past IP address tasks, so it can specially allocate to a customer a similar IP address that the customer recently had. The DHCP worker gives a private IP address subordinate upon each's customer id dependent on predefined planning by the overseer. This element is differently called static DHCP task by DD-WRT, fixed-address by the DHCP documentation, address reservation by Netgear, DHCP reservation, or static DHCP by Cisco and Linksys, and IP address reservation or MAC/IP address authoritative by different other switch makers.

4. DNA Cryptography.

Deoxyribonucleic Acid (DNA) cryptography [8], the utilization of the qualities of DNA offers new chances and heading to information stowing away. This work will use the natural qualities of DNA successions. The instruments of DNA capacity, for example, integral blending, and DNA record give another layer of security to the proposed technique.

So as to secure delicate information through unstable systems like the Internet, utilizing different sorts of information insurance is fundamental. One of the popular approaches to secure information through the Internet is information covering up. In view of the expanding number of Internet clients, using information stowing away or Steganographic methods [9] is unavoidable. Wiping out the job of the gatecrasher and approving the recipient, are possible objectives of these strategies. In this way, the job of information covering up has become more prominent these days. Before utilizing organic properties of DNA arrangements, normally implanting a mystery message into the host pictures was the conventional method of information covering up. Tragically, this had a few liabilities. The most significant ones were the recognition of the mutilations of the picture when the host picture changed to certain degrees. This spot was the best spot to begin the complete recognition of the mystery message through the picture.

By coming of organic parts of DNA arrangements to the registering regions, new information concealing strategies have been proposed by scientists, in view of DNA successions. The key bit of their work is, using organic attributes of DNA arrangements.

5. Review of Literature:

Salah Alabady et al [10] is executed a Network Security Model for Cooperative Network introduced a system security model. The creator has examined weaknesses, dangers, assaults, arrangement shortcomings, and security strategy with system assurance.

Balaraju. J et al [11] have examined big data advances and their advances for expanded large information. Information security is a chief issue in the administration part, science, exploration, and business ventures. They likewise examined information stockpiling, handling, and security territory and discover the challenges by utilizing regular security instruments for Hadoop. At last, the Authors supported a validation component utilizing complex DNA cryptography an answer for huge information security as opposed to changing over large information into DNA with less computational assignments. They recommended a solitary DNA based secure hub for verification and metadata the board for Hadoop which is the best answer for strengthening information and dispose of NNSE blocks for security metadata in the Namenode in Hadoop.

Mohammed Nadir Bin Ali et al [12] developed a Secure Campus Network have configured. The developed hierarchical architecture of the campus network considering different types of security issues that ensure the quality of service.

Kartik Pandya et al [13] are developed a Network Structure and discussed five basic network topologies like Bus, ring, Star, Tree, and Mesh.

Balaraju.J et al [14] are built up an Algorithm Built-in Authentication Based on Access (BABA) as a security occurrence coordinated as Hadoop hub for making sure about information in HDFS and promptly metadata security for evading clients information in Hadoop. The instrument contributes a made sure about Hadoop Cluster without utilizing other security game plans which likewise lessens operational cost, calculations, expands information security, and giving stable security answers for Hadoop Cluster. By utilizing this, there is a degree for arranging Single Node Clusters in the association by lessening operational, computational expense, and expanding information security, and gives better security to MNC's. The improvement of this work is to lessen the computational weights of the proposed calculation.

Kennedy et al [15] are executed a Structured Network for a Small system. Creators have reenacted organize configuration utilizing the Cisco Packet Tracer programming and Wire Shark convention analyzer.

Balaraju.J et al [16] had executed an only new security component, a Secure Authentication Interface (SAI) layer over the Hadoop Cluster. As a Single Security convention, this interface presents client verification, metadata security, and access control. Contrasted with the current instruments, SAI can give security a less computational weight. Creators concentrated on security challenges and tended to for making sure about Big Data in a Hadoop Cluster through a solitary, restrictive security system called Secure Authentication Interface. SAI made a confided in condition inside HC by confirming the two clients and their cycles.

All the above papers that are surveyed have proposed various parts of distributed system structure, geographies, and execution yet they have not examined issues looked in viable usage. Numerous Authors are concentrating a made sure about the appropriation framework in Hadoop and eventually which is valuable to information in the Hadoop group have not examined DHCP and MAC officials in detail.

6. Problem Statement.

In any distribution system including Hadoop distribution systems connected a centralized network switch. Each node in the network has a unique IP address that is dynamically assigned by the DHCP server by collecting the physical address of each connected Node and assigning static IP difficult to network admin in larger networks. IP address and hostname can appear any user working in a distributed system and they may access any node with appropriate permissions. The problem with appearing IP and host, the hacker may place duplicate hostname and duplicate IP address for disturbing or malfunctioning network by placing a malicious node in the network. Since there is scope losing important data from a distributed system by using IP address and it can be a security threat. This threat can be reduced by hiding the IP address and hostname without appearing any user working in a distributed environment.

7. Related Work.

The present invention relates to distributed computing systems and is more particularly directed to architecture and implementation of a scalable distributed computing environment which facilitates communication between independently operating nodes on a single network. The primary objective of the research work is to create a layer on top of the distributed system especially for the Hadoop distributed system, so every node appears in the layer. For setting the environment we configured a centralized DHCP server and 200 nodes in the network with a different configuration. Multinode HC[17] configured, the master node configured within the DHCP server as Namenode, and the remaining nodes are data nodes. The developed security layer is also configured within the DHCP Server for collecting the hostname and IP addresses which is stored itself.

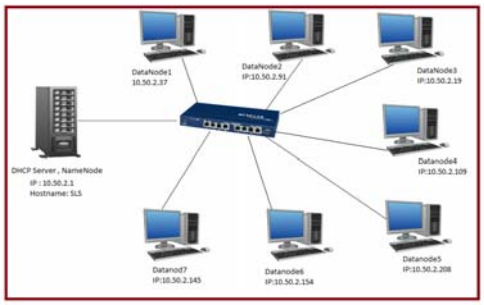

Figure 1: Existing Hadoop Cluster by configuring DHCP Server.

The DNA algorithm within the proposed layer is converting the IP address, the hostname in different level and producing a unique key which appears for the user including the physical address (MAC) of NIC. The key generation is generating by using a highly secured DNA hiding [18] methodology for creating confusion hackers to access any node from the network and finally it becomes a highly secured distributed system.

Algorithm For generating Unique Key for Node:

Start :

1. *Collecting IP,Hostname,MAC*
2. *Combining IP,Hostname,MAC as UNIQID*
3. *Converting UNIQID into BINARY Form*
4. *Converting Binary form to DNA*
5. *Assign a num toDNA form and forming a Decimal number // a=0,c=1,g=2,t=3.*
6. *Converting Decimal number to Hexadecimal.*
7. *Generating UNIQID from Hexadecimal.*

Stop

Table 1- Procedure for hiding system information using DNA.

Steps	Description	MAC_Address	IP Address	Hostname
01	Node Data	10:78:D2:55:95:A8	10.50.4.8	HD_DN01
02	Combined Node Info.	10:78:D2:55:95:A8-10.50.4.8- HD_DN01		
03	Binary form	00110001 00110000 00111010 00110111 00111000 00111010 01000100 00110010 00111010 00110101 00110101 00111010 00111001 00110101 00111010 01000001 00111000 00101101 00110001 00110000 00101110 00110101 00110000 00101110 00110100 00101110 00111000 00101101 00100000 01001000 01000100 01011111 01000100 01001110 00110000 00110001		
04	DNA Form	ATACATAAATGCATGTATGAATGGCA CAATAGATGGATCCATCCATGGATGC ATCCATGGCAACATGAAGTCATACAT AAAGAGATCCATAAAGAGATCAAGA GATGAAGACAGAACAGACACACCTT CACACATGATAAATAC		
05	A=0, C=1, G=2, T=3	0301 0300 0321 0323 0320 0322 1010 0302 0322 0311 0311 0322 0321 0311 0322 1001 0320 0231 0301 0300 0202 0311 0300 0202 0310 0202 0320 0201 0200 1020 1010 1133 1010 1030 0300 0301		
06	Decimal - Hexa	5CC3148053C4906F901A54B4DE8225FD 8071B87BA02E51C8E7B2C2BDE75CC31 47A2CCCD09B85CEA1AB7E9E03DF1AE C6D5947237BCFA358FB2DED		
05	Unique Key	5CC3148053C4906F901A54B4DE8225FD 8071B87BA02E51C8E7B2C2BDE75CC31 47A2CCCD09B85CEA1AB7E9E03DF1AE C6D5947237BCFA358FB2DED		

The hostname of the node is 8 characters, it dynamically assigning from generated unique key and changing every 7 days, it can be updated automatically central table. So, the hackers get confusions to access a particular hostname from the distributed system. The main advantage of the proposed method is not having permanent identification like hostname to access node and it shared data.

The Secured layer also contains nodes information by maintaining central table along with hostname. Internally every node connected with other nodes by appearing hostname controlled by the security layer. All these hostnames in the network are maintained by a

secure layer including the storage status, processing configuration by periodically updating in the central table. The secure layer is updating the central table when a new node is added or removing of a node from the distributed system.

Table 2 – Nodes Status information in Central Server.

MAC	Node_Hostname	Status	Joined / Removed
A4-1F-72-58-BB-01	A1AB7E9E0	Active	14-MAR-2020
F5-10-72-58-BB-01	3DF1AEC6D	Active	14-MAR-2020
C4-1F-B2-58-BB-01	5947237BC	Active	14-MAR-2020
A4-1F-72-58-BB-01	FA358FB2D	Removed	23-JUN-2020
A4-1F-72-58-BB-01	47237BCFA	Active	14-MAR-2020
A4-1F-72-58-BB-01	CC3147A2C	Active	14-MAR-2020
A4-1F-72-58-BB-01	BDE75CC31	Active	14-MAR-2020
A4-1F-72-58-BB-01	2E51C8E7B	Removed	05-JUL-2020

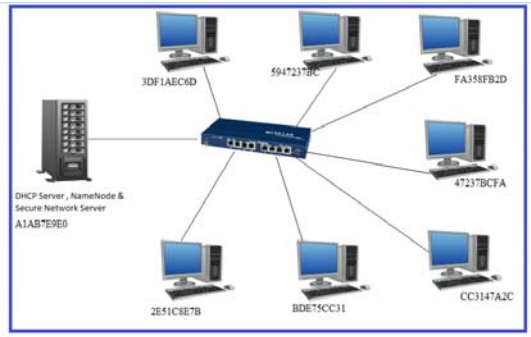

Figure 2- Proposed Hadoop Cluster by Dynamic Hostnames.

8. Performance Evaluation.

Performance analysis is analyzed with small distributed system with proposed secure layer All the existing security methods are not concentrated to hide nodes information.

Table 3 - Nodes and Users Information

	Existing	Proposed (Security Layer)
Nodes	250	250
Users	207	252
Nodes Accessed	225	9
Results (data Access Nodes)	180	0

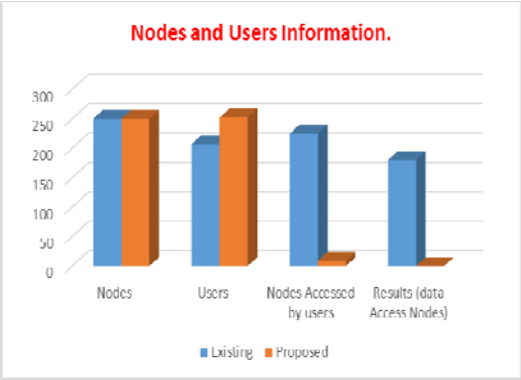

Figure-3. Nodes and User Data Existing and Proposed

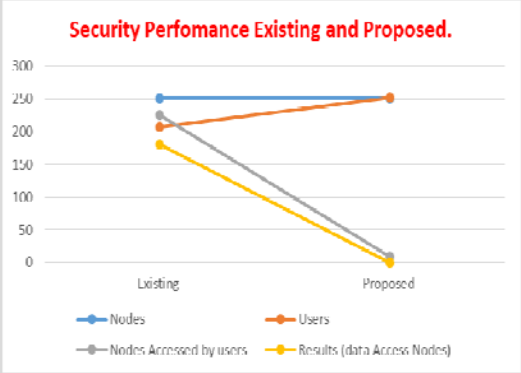

Figure-4. Performance evolution of Existing and Proposed

This security layer likely provides 24X7 security for HC which is very useful for small distributed system to maintain their data securely. This can increase the data security, communication operational, and reduce maintenance problem

9. Conclusion.

In the current research work, we have implemented a secured distributing system with the help of DHCP server IP, HOST, and MAC combination. The distinctive distributed system network allocated a dedicated UNIQID to each node for securing Network. The administrator only has complete privileges to access all the nodes including the server and the others cannot access any node in the network without knowing IP or Hostname. The performance of the network and distributed system and network is increased ultimately data security is enhanced. The entire network is planned to be automatic which involves minimum user intervention. The future scope of this work is to enhance

the same security layer to apply to the WAN network located in different locations.

References.

[1] H. Yang and J. Lee, "Secure Distributed Computing With Straggling Servers Using Polynomial Codes," in IEEE Transactions on Information Forensics and Security, vol. 14, no. 1, pp. 141-150, Jan. 2019, doi: 10.1109/TIFS.2018.2846601.

[2] Khanan A., Abdullah S., Mohamed A.H.H.M., Mehmood A., Ariffin K.A.Z. (2019) Big Data Security and Privacy Concerns: A Review. In: Al-Masri A., Curran K.(eds) Smart Technologies and Innovation for a Sustainable Future. Advances in Science, Technology & Innovation (IEREK Interdisciplinary Series for Sustainable Development). Springer, Cham. https://doi.org/10.1007/978-3-030-01659-3_8.

[3] Roy M. et al. (2020) Data Security Techniques Based on DNA Encryption. In: Chakraborty M., Chakrabarti S., Balas V. (eds) Proceedings of International Ethical Hacking Conference 2019. eHaCON 2019. Advances in Intelligent Systems and Computing, vol 1065. Springer, Singapore. https://doi.org/10.1007/978-981-15-0361-0_19.

[4] R. Samet, A. Aydın and F. Toy, "Big Data Security Problem Based on Hadoop Framework," 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 2019, pp. 1-6, doi: 10.1109/UBMK.2019.8907074.

[5] Akhgarnush E., Broeckers L., Jakoby T. (2019) Hadoop: A Standard Framework for Computer Cluster. In: Liermann V., Stegmann C. (eds) The Impact of Digital Transformation and FinTech on the Finance Professional. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-23719-6_18.

[6] A. K. Rajput, R. Tewani and A. Dubey, "The helping protocol “DHCP”," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 634-637.

[7] M. Mohsin and R. Prakash, "IP address assignment in a mobile ad hoc network," MILCOM 2002. Proceedings, Anaheim, CA, USA, 2002, pp. 856-861 vol.2, doi: 10.1109/MILCOM.2002.1179586.

[8] K. S. Sajisha and S. Mathew, "An encryption based on DNA cryptography and steganography," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 162-167, doi: 10.1109/ICECA.2017.8212786.

[9] Nirmalya Kar, Kaushik Mandal, Baby Bhattacharya, Improved chaos-based video steganography using DNA alphabets, ICT Express, Volume 4, Issue1, 2018, Pages 6-13, ISSN 2405-9595, https://doi.org/10.1016/j.icte.2018.01.003.

[10] Salah Alabady, Design and Implementation of a Network Security Model for Cooperative Network, International Arab Journal of Technology, Vol. 1, No. 2, June 2009.

[11] Balaraju, J. and P. V. V. P. Rao. “Recent advances in big data storage and security schemas of HDFS: a survey.” (2018).

[12] Mohammed Nadir Bin Ali, Mohamed Emran Hossain, Md. Masud Parvez. Design and Implementation of a Secure Campus Network International Journal of Emerging Technology and Advanced Engineering Website, 2015, 5(7). www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal.

[13] Kartik Pandya. 2013, 1(2). 6. International Journal of Advance Research in Computer Science and Management Studies Network Structure or Topology.

[14] Balaraju, J., Prasada Rao, P. V. R. D.: Designing authentication for Hadoop Cluster using DNA algorithm. Int. J. Recent. Technol. Eng. (IJRTE) **8**(3) (2019). ISSN: 2277-3878. https://doi.org/10.35940/ijrte.C5895.0983.

[15] Offor, Kennedy J, Obi, Patrick I, Nwadike Kenny T, Okonkwo II. International Journal of Engineering Research & Technology (IJERT), 2013, 2(8).

[16] Balaraju J., Prasada Rao P.V.R.D. (2020) Innovative Secure Authentication Interface for Hadoop Cluster Using DNA Cryptography: A Practical Study. In: Reddy V., Prasad V., Wang J., Reddy K. (eds) Soft Computing and Signal Processing. ICSCSP 2019. Advances in Intelligent Systems and Computing, vol 1118. Springer, Singapore. https://doi.org/10.1007/978-981-15-2475-2_3.

[17] Gugnani S., Khanolkar D., Bihany T., Khadilkar N. (2014) Rule Based Classification on a Multi Node Scalable Hadoop Cluster. In: Fortino G., Di Fatta G., Li W., Ochoa S., Cuzzocrea A., Pathan M. (eds) Internet and Distributed Computing Systems. IDCS 2014. Lecture Notes in Computer Science, vol 8729. Springer, Cham. https://doi.org/10.1007/978-3-319-11692-1_15.

[18] Demidov V.V. (2020) Hiding and Storing Messages and Data in DNA. In: DNA beyond Genes. Springer, Cham. https://doi.org/10.1007/978-3-030-36434-2_2.

Mr.J.Balaraju. is Research Scholar in Computer Science & Engineering department at Koneru Lakshmaiah Educational Foundation (Deemed To be University), Vijayawada. Working as an Assistant Professor in the Computer Science & Engineering Department at Rajeev Gandhi Memorial College of Engineering & Technology (Autonomous), Nandyal, AP, - India and His research areas include Big Data Analytics, Data Mining, IoT and Sensor network..

Dr.PVRD.Prasada Rao is a Professor in the Department Computer Science & Engineering at Koneru Lakshmaiah Educational Foundation (Deemed To be University), Vijayawada. His research areas include data mining, bioinformatics, IoT, Sensor network and big data analytics. He has published 70+ research papers in the leading international journals and conference proceedings. In addition he is an Associate Dean (P&P) / Reviewer/ Member of several international conferences / workshops.