Int. J. Advance Soft Compu. Appl, Vol. 12, No. 3, November 2020
ISSN 2074-8523; Copyright © ICSRS Publication, 2020
www.i-csrs.org

Investigation and Finding A DNA Cryptography layer for Securing data in Hadoop Cluster

Balaraju J.[1], PVRD. Prasada Rao[2]

[1]Research Scholar, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
e-mail: jb7443@gmail.com
[2]Professor, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
e-mail: pvrdprasad@kluniversity.in

Received 28 June 2020; Accepted 10 October 2020

Abstract

Big data is essential for the modern digital world and many IT industries, researchers, data specialists are using it for storage and analysis. The most significant issues are, storage, analysis, and securing data. Hadoop is a very versatile, adaptable platform for storage of data and focusing data analytics, data driven applications, however it was not begun in light of security or authorization for data. Security and controlling data are most significant fragments of any distributed platform that wants to break into the endeavor standard. Hadoop is fitting for dealing with efficient in storage and analysis of data and it has certain security issues and also it uses third party security which used huge computations in all the versions of Hadoop. This paper proposed a new security mechanism as Secure Hadoop Cluster using DNA Cryptography (SHCDNA) as single security instance gathering metadata from Namenode at regular intervals. The SHCDNA is a single security instance for user authentication and it maintains user sensitive information in an encrypted form and metadata of users. The proposed security is to offer a better solution by improving security and performance by reducing the computations of existing Hadoop security mechanisms.

Keywords: *Big data, Security, Authentication, Cryptography, SHCDNA.*

1 Introduction.

Big data [1] is playing an important role in the modern world and a huge volume of data is generating from different sources like social media, sensors, retail logistics, and the Internet of things. Traditional distributed technologies are not suitable for storage and analysis of big data because the data is unstructured, semi-structured, and structured in nature. In the digital world, big data is not only for storage of petabytes and Exabyte's of data in clouds and also a need for provide data available to business organizations.

2 Data Security Requirements.

Predominantly, Data security requires mainly three properties that are data integrity, confidentiality and availability. The Integrity handles protecting data from unofficial changes. Privacy or Confidentiality refers to securing data from illegal data accesses. Data respectability has been additionally summed up to information reliability which implies data isn't just changed by unapproved subjects yet in addition it is liberated from mistakes and modern just as beginning from trustworthy sources. However, the data constancy is the major difficult. The solution wants merging different methods ranging from cryptographic procedures for digitally validation providing on data access control, for scrutiny that only legal users to modify data, to data quality systems, for repeatedly finding and fixing data errors, provenance procedures, for determining from which bases data patent and repute methods for measuring the reputation of data sources. Finally, accessibility is the property of guaranteeing that information are accessible to approved clients. These three prerequisites are extremely basic still today and contacting them is likewise considerably more testing since information assaults are more advanced and the information assault surface has extended, because of expanding information assortment exercises from various sources and to information sharing.

3 Hadoop.

Open source Application like Hadoop [2], NoSQL [3] databases and cloud computing are available to handle big data. The core part of Hadoop for massive data storage called Hadoop Distributed File System (HDFS) [4] and processing techniques called Map Reduce [5]. Hadoop is a fully distributed platform by supporting low and high level configurations with huge data analytics, data-centric systems. In any hadoop release, it was not considered in light of security or authority of data in all versions. Security and control of data are crucial mechanisms of any data distribution that hopes to interruption into the enterprises.

3.1 Hadoop Security.

Hadoop has scalable distribution in different releases, bendy Bigdata era assisting all the ways of information processing workloads and focused analytics and data centric applications. Security used to be now not the pinnacle of thought when Hadoop was once at the beginning developed. There have been countless motives for this, inclusive of that all the statistics being listed by using Yahoo! used to be open source and it was only one application system access through a trusted users. From the beginning, Hadoop weak in its fundamental security abilities like user verification and data access controls. As Hadoop itself developed to manage new sorts of outstanding tasks at hand and applications, many including sensitive data, a few wellbeing capacities have been created through each the open-source and seller networks. Unfortunately, these security abilities have been commonly evolved in confinement from each other. This is in phase due to the fact Hadoop not have a single security mechanism, however a series of open projects, sub-projects and enterprise application extensions which are work collectively in quite a number mixtures to allow extraordinary patterns of data-centric applications.

4 DNA Cryptography

Many researchers are focused on using DNA [6] security concepts to improve the security of handling big data by the popular HDFS. Genetic data is coded as a structure of nucleotides like Cytosine, Adenine, Guanine and Thymine. A DNA sequence [7] consists of four 4 letters of the alphabet by representing binary digits 00, 01,10, and 11 for A, C, G, and T. DNA cryptography [8] methods are the most prominent and captivating for Big data security as of the multipart structure of DNA. The most trivial future is no direct association DNA with information security [9]. Big data security using DNA is in a preliminary stage, many researchers are focused to develop DNA based security protocols.

5 Literature Survey on DNA Based Security Approaches.

Balaraju and PVRD Prasada Rao et al [10] had analyzed big data technologies and their advances for increased big data. Data security is a foremost issue in the government sector, science, research, and business enterprises. They also analyzed data storage, processing, and security area and find out the difficulties by using conventional security tools for Hadoop. Finally, the Authors endorsed an authentication mechanism using complex DNA cryptography a solution for big data security instead of converting big data into DNA with less computational tasks. They suggested a single DNA based secure node for authentication and metadata management for Hadoop which is the best solution for fortifying data and discard NNSE hindrances for security metadata in the Namenode in Hadoop.
Suyel Namsudra et al [11] had investigated and developed a data access control model using DNA for cloud data centers based on the size of data shared by data

owners. They used a 512-bit secret key by using users' attributes or decryption key for data storage. The model provides cloud services and the users can pay fewer amounts for utilizing cost. The main objective is to provide efficient data security, reducing data access time and efficient data storage in the cloud.

Balaraju and PVRD Prasada Rao et al [12] has designed developed an DNA data hiding algorithm Built-in Authentication Based on Access (BABA) and it is a security occurrence combined as a Hadoop node data in HDFS and instantaneously metadata security for dodging users data in Hadoop. The instance tool contributes a secured Hadoop Cluster without configuring other security tools which also reduces operative cost, computations, data security increases and providing stable security solution for Hadoop Cluster. By using this, there is a scope for configuring Single Node Clusters in the organization by reducing operational, computational cost and increasing data security and provides better security for MNC's. The enhancement of this work is to reduce the computational burdens of the proposed algorithm.

Balaraju and PVRD Prasada Rao et al [13] has implemented an exclusively new way security system, a Secure Authentication Interface (SAI) as a secure layer which positioned above the Hadoop Cluster. As a Single Security protocol, this interface presents user verification, protecting metadata, and data access control. This developed protocol SAI is providing security using fewer computations. The Authors focused on security challenges and addressed for securing Big Data in open source Hadoop using a standard and developed exclusive security protocol called Secure Authentication Interface. This security systems created a trusted background for HC by validating both users and its jobs.

6 Problem Definition

Integrating a third-party authentication protocol like Kerberos [14] in the Hadoop Cluster is quite tedious task. HC required a single and own efficient security instance for Granting and Revoking capabilities. Many Hadoop distributions are configuring separate security protocols like NNSE [15] is for metadata security collecting from Name Node. Authentication is essential and plays a crucial role in securing data, all the distribution is using Kerberos authentication and it is not proprietary of Hadoop which is using high computational time. For securing sensitive data bull's eye algorithm is configuring Hadoop based distributions. A single security protocol is mandatory for HC which is suitable for Authentication, metadata, and securing the sensitive data.

7 Related Work.

The Datacenter is configuring with Hadoop and a Secure SHCDNA node for authentication and managing metadata instead of configuring two separate

security like Kerberos and NNSE. The performance of Hadoop data centers is not up to the mark due to the above security mechanisms which are proprietary. The proposed method paves the way for better security and improves performance strong authentication and managing metadata with fewer computations.

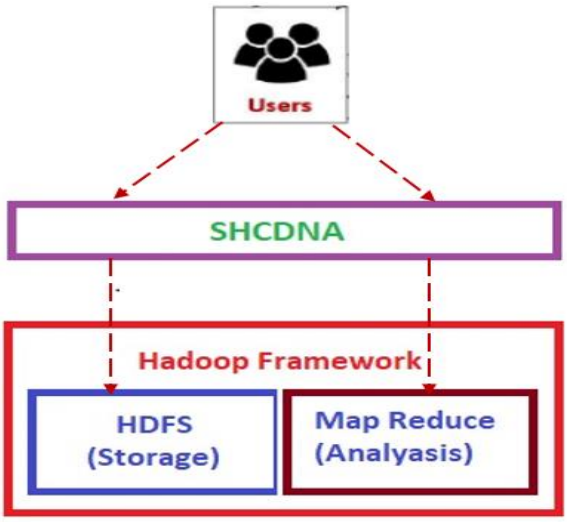

Fig-1. Proposed Secure Framework with SHCDNA.

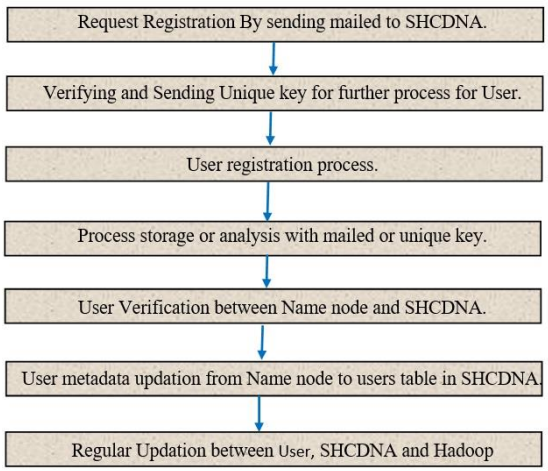

Fig-2. Steps for Users, SHCDNA and Hadoop framework.

7.1 Working with SHCDNA.

SHCDNA is playing a vital role in authenticating users in different phases like data storage and analysis in HC. Every new user must send a request to SHCDNA by sending an email id and it is generating a permanent unique key for each user using DNA.

7.2 User unique ID's using DNA.

SHCDNA forwards the created unique key to the user for the further registration process, which includes name, mobile number, address, and other credentials. The user information is to be updated in the user table in the SHCDNA node by verifying the user unique key. The user also submits mail id and unique key to the NameNode for availability the same in the SHCDNA for the confirmation. All the user activities and dynamic keys are regularly updating in the user's table in SHCDNA. Users' authorization is and approvals are completely handled by SHCDNA with the help of NameNode at regular intervals. Once the complete process is completed, the SHCDNA is establishing a trusted environment user and Hadoop.

Table-1- Unique Key and User information Table.

Mailid	Unique Key	Metadata	Name	Counter
bds@gmail.cm	98-100-115-64-103-109-97-105-108-46-99-111-110	DC1, DC2, DC5	Balaraju	1

Once the user stored data in Hadoop the NameNode having corresponding users metadata and it can be updated regularly every 8 seconds. The SHCDNA is also acquiring metadata from NameNode in regular intervals for managing users Metadata. NameNode verifies the unique key and sends the corresponding Metadata to SHCDNA and it will update the metadata in the central table in every 8 seconds. The users can access their data from DataNodes by acquiring metadata from the central table of SHCDNA even in the absence of NameNode and Secondary Name Node. This can be used as the best solution for Hadoop.

8 Experimental Results for Metadata Security.

Table-2. Metadata Security existing and proposed Mechanisms.

	Number of times Updation Required(Existing)	Number of times Updation Required (Proposed)
User1	7.5	1
User2	18.75	2.5
User3	25	3.18
User4	75	10
User5	37.5	5.1

The above table is showing for 5 Hadoop users metadata information existing which is updating every 8 seconds, but the proposed security mechanism can update every 60 seconds for reducing the computational burdens.

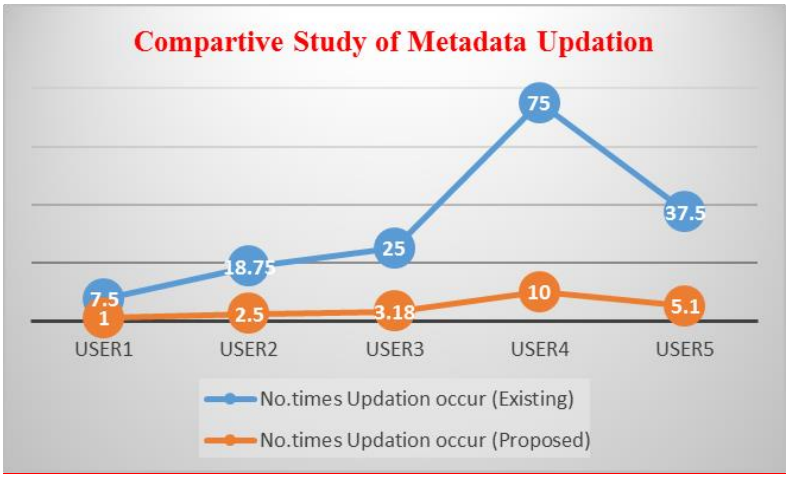

Fig-2. Comparative study Metadata updation in existing and proposed

8.1 Comparative study of Results 1.

Table-3- No.of Security Configurations existing and proposed

	Existing	Proposed
No. of Security Configuration	02	01
No. Computation for 1 user Authentication	10	04
Metadata security Computations per minute	9	00

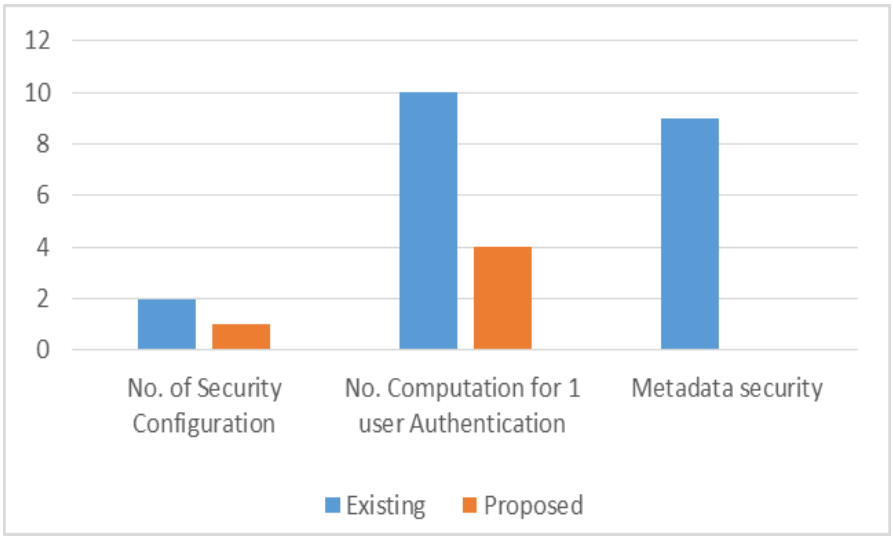

Fig-3. Security configuration, computational time for existing and proposed.

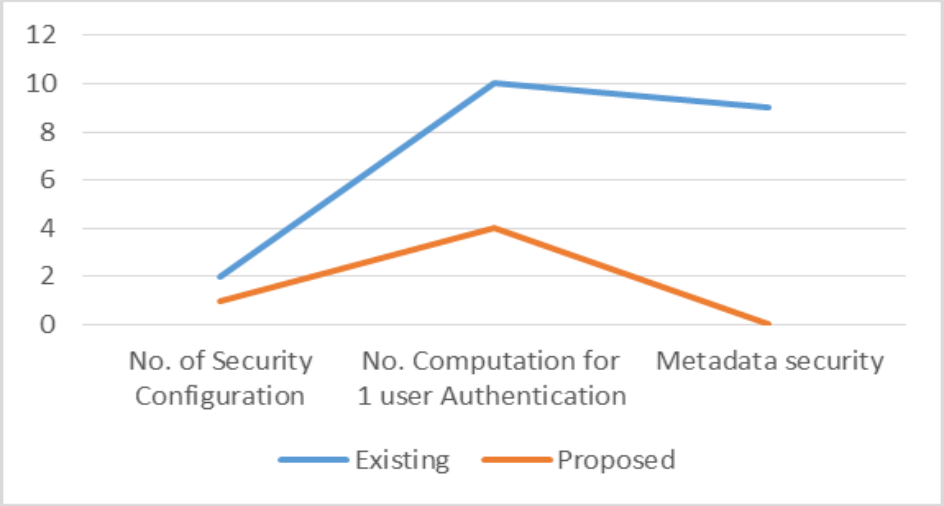

Fig-4. Performance of existing and proposed mechanisms.

The following table showing the 10 users wants login single node cluster and the required to authentication by using Kerberos before entering the cluster. Every time they request KDC for each session which is taking 10 computation per session, each users may request authentication KDC depending on their usability of cluster.

Table-3. User Authentication computation existing and proposed.

	No.of sessions Authentications /Day (Existing)	Total Computations	Number of Time Authenticated / Day(Proposed)	Total Computations
User1	5	50	2	8
User2	**10**	**100**	**3**	**12**
User3	8	80	4	16
User4	6	60	1	4
User5	2	20	2	8
User6	3	30	1	4
User7	7	70	3	12
User8	**6**	**60**	**1**	**4**
User9	3	30	2	8
User10	2	20	1	4
Total	**52**	**520**	**20**	**80**

Table 3 shows the user sessions and total computations of existing and proposed authentication.

For existing authentication User2 has used highest authenticated sessions per day of 10 compared to other users based on accessing of HDFS data. The existing authentication used average computations of 10 for 10 users.

The proposed authentication mechanism User3 has used highest authenticated sessions per day of 4 compared to other users. The proposed authentication used reduced average computations of 4 for 10 users with more data security.

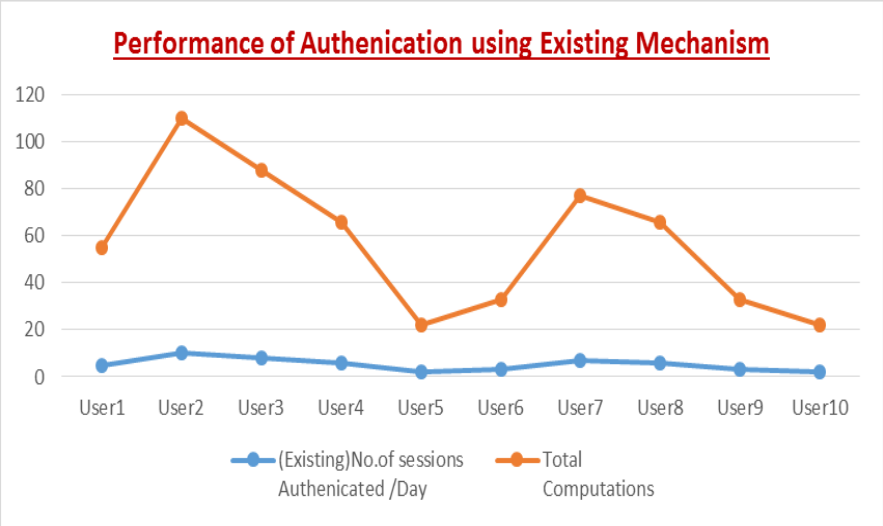

Fig-7. Computational Performance of existing and proposed mechanisms.

9 Conclusion and Future Scope.

Hadoop uses a different Distributed environment compared to conventional Distributed System. The Kerberos authentication protocol is developed based on a distributed system that required more computations and it completely depends on Kerberos Distribution Center. HC uses Kerberos as one of the core authentication security protocols. HC users cannot read their data in the absenteeism of master NameNode and it required secondary NameNode, based on this the performance of HC bottleneck. To overcome this, HC needs its protocol SHCDNA which serves the purpose of authentication and metadata security. The vision of this developed system is to increase data access in the absence of Namenode which ultimately improves performance and Data Security. The future work is metadata encryption using DNA.

References

[1] Furht B., Villanustre F. (2016) Introduction to Big Data. In: Big Data Technologies and Applications. Springer, Cham.

[2] Xu H., Chen X., Fan G. (2020) Ecosystem Description of Hadoop Platform Based on HDFS, MapReduce and Data Warehouse Tool Hive. In: Xu Z., Choo KK., Dehghantanha A., Parizi R., Hammoudeh M. (eds) Cyber Security Intelligence and Analytics. CSIA 2019. Advances in Intelligent Systems and Computing, vol 928. Springer, Cham.

[3] Chen J-K, Lee W-Z. An Introduction of NoSQL Databases Based on Their Categories and Application Industries. Algorithms. 2019; 12(5):106.

[4] Wei Dai, Ibrahim Ibrahim, Mostafa Bassiouni, "An Improved Replica Placement Policy for Hadoop Distributed File System Running on Cloud Platforms", Cyber Security and Cloud Computing (CSCloud) 2017 IEEE 4th International Conference on, pp. 270-275, 2017.

[5] Huang, W., Wang, H., Zhang, Y. et al. A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop. Cluster Comput 22, 13077–13084 (2019). https://doi. org/10.1007/ s10586-017-1205-9.

[6] Lee, S., Lee, E., Hwang, W. et al. Reversible DNA data hiding using multiple difference expansions for DNA authentication and storage. Multimed Tools Appl 77, 19499–19526 (2018). https://doi .org/10.1007/s11042-017-5379-1.

[7] He PA, Wang J. , "Characteristic sequences for DNA primary sequence." , J Chem Inf Comput Sci. 2002 Sep-Oct;42(5):1080-5.

[8] A. Vikram, S. Kalaivani and G. Gopinath, "A Novel Encryption Algorithm based on DNA Cryptography," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019, pp. 1004-1009, doi: 10.1109/ICCES45898. 2019.9002399.

[9] 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.

[10] Balaraju, J. and PVRD Prasada Rao. "Recent advances in data storage and security schemas of HDFS: a survey." (2018).Special Issue (Emerging Trends in Engineering Technology) March-2018, PP. 132-138, SCIE indexed.

[11] Suyel Namasudra, Pinki Roy, Pandi Vijayakumar, Sivaraman Audithan, Balamurugan Balusamy, Time efficient secure DNA based access control model for cloud computing environment, Future Generation Computer Systems, Volume 73, 2017, Pages 90-105, ISSN 0167-739X.

[12] 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.

[13] 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.

[14] Nan Zhang, Xiaoyu Wu, Cheng Yang, Yinghua Shen and Yingye Cheng, "A lightweight authentication and authorization solution based on Kerberos," 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Xi'an, 2016, pp. 742-746,doi:10.1109/IMCEC.2016. 7867308.

[15] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan,Big Data and Hadoop-a Study in Security Perspective, Procedia Computer Science, Volume 50, 2015, Pages 596-601, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.091.

Notes on contributors

Mr. J.Balaraju. is Research Scholar in Department of Computer Science & Engineering at Koneru Lakshmaiah Educational Foundation (Deemed To be University), Vijayawada. And Working as a 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.