


Ph.D. Research Proposal

Doctoral Program in “Department Name”

Secure Data Deduplication and Classification in Hadoop

Distributed File System using Intelligent Solutions

 **by**
PHD PRIME
YOUR RESEARCH PARTNER
<Name of the Candidate>

<Reg. No of the Candidate>

<Supervisor Name>

<Date of Submission (DD MM 20YY)>

I. INTRODUCTION / BACKGROUND

Security in big data is an emerging topic which is considerable interest among researchers since deployment of big data storage systems are valuable target by intruders. “Big Data technologies define a new generation of architectures, technologies designed to capture, store, update, manage, and analyse of large volumes of a wide variety of data by enabling high-velocity capture, discovery and or analysis. Loosely speaking, big data sets are diverse, large and complex generated from sensors, instruments, email, video, and instruments”.

With the use of huge volume of data, big data storage systems lead to big complexity and also security breaches occur, which is difficult to solve in real-world environment. However, cloud server or any big data server is not trustworthy. For example, healthcare applications require secure large storage server since genome data must be stored securely and it is size of 140 gigabytes. Due to this security breaches, data owner disclose their private and sensitive data to the cloud or big data server.

Various security attacks occur in big data storage systems such type of attacks are Password Guessing Attack, Brute force Attack, Stolen Verifier Attack etc. [8], [11]. Existing security approaches have proposed to secure data by sending data in a form of cipher text. But, these approaches are failed to provide confidential and privacy for data owners and users. For now data security is a big concern in big data. From security perspective it is crucial due to:

- Access policy is not designated when ciphertext is updated and here user legitimacy is failed that means who intends to access the data are still great concern in big data
- There is no authorized entity to monitor the data sharing and outsourcing to storage systems.

For big data storage system, security (authentication, data confidentiality and integrity) monitoring is vital in real-time. Some of real-time applications are smart grid, transaction application and e-healthcare applications [10], [12]. To ensure authorization and data security,

the main solution is encryption of data, which helps in provide adequate protection of encrypted data [1], [4]. Similarly aggregation of large volume of data requires less storage space. Big data security with clustering of variety of data is the great aspect to be considered. However, big data clustering may cause serious problems such as data loss [2]. This is mainly due to the encryption after compression of data [3]. Compression before encryption can easily handle such errors in compression. Henceforth, many secure clustering approaches have been proposed. For clustering similarity based approach is proposed which is called Locality Sensitive Hashing (LSH) [5]. It is a similarity approach to compute the similarity between two datasets. Hierarchical attribute based encryption is proposed [6], [7]. LSH is only suitable for small size of data and it is not suitable to find similarity between large size of data, particularly in big data (terabytes, gigabytes and petabytes of data).

1.1 Research Outline & Scope

The main aim and scope of this research work is to provide secure environment for data deduplication and classification by distributed and intelligent solutions.

1.2 Research Objectives

To provide best secure environment for both data users and owners in big data assisted cloud servers. In this work, reduce the response time for data users when request for data to retrieve.

II. RESEARCH GAPS

2.1 Common Problem Statement

For big data storage system, security requirements (authentication, data confidentiality and integrity) monitoring is vital in real-time. To ensure authorization and data security, the main solution is encryption of data, which helps in provide adequate protection of encrypted data.

2.2 Problem Definition

In this paper, access policy created [1] by data owner is updated in the cloud and also updating a cipher text securely. To update a new access policy, in this paper authors have

proposed a secure and verifiable access control scheme using improved NTRU cryptosystem. This paper based on the big data storage in cloud environment. However, this NTRU cryptosystem has failed by two decryption failures such as Wrap failure and Grap failure. Furthermore, secret sharing is presented based on (t, n) thresholding.

Problem

- NTRU is very strong secure and fast public key cryptosystems than RSA, ECC. The parameters of NTRU are private key, public key, encryption and decryption. The limitation of this NTRU is it can be failed while encrypting large amount of data.
- Cloud access control system can be hacked by attackers. When data owner is hacked, attacker has full rights to access private information. Attacker can also use that to get through other access control legitimately. Hence trusted third party is required to manage and control data owners.

Proposed Solutions

- To encrypt large volume of data, we proposed AES (128bits, and 256bits). Data classified by MapReduce is the distributed data processing framework which mainly created to process large volume of data
- We used Trusted Center to register and monitor data owner's behavior. Access control is given to all encrypted data since data owners encrypt data based on the sensitivity level. Data owners classify file into any of the categories according to events. For these three categories three type of encryption procedure is presented.

This paper addressed two problems [2] include clustering and data security in order to prevent from data loss. Hence in this paper, a novel multidimensional clustering scheme is proposed which eliminate data loss and protect data from security attacks by SDES (Simple Data Encryption Standard) encryption technique. In addition, Huffman compression is introduced which control the data size and prevents from large overheads. The flow of this work is following: (1). Get the big text file as input, (2). Implement SDES algorithm on input file for encryption, (3). Apply Huffman compression technique of encrypted data, (4). Error control

technique is applied on compressed file for error correction and (5). Apply clustering on error controlled data.

Problem

- However, compress the input file and then encryption is the best way. This work failed to increase the computation time since when first encrypt data and then compress, we would no gain performance improvement related to speed. It is only suitable for less data
- SDES encryption algorithm is very simple and does not provide high level security

Proposed Solutions

- We firstly recommend data compression and then perform data encryption, which reduce the problems of previous works. We propose deduplication by Jaccard Similarity c , which is best than others
- AES Encryption algorithm is proposed to mitigate the issues of SDES encryption, which is more secure and adequate for large volume of data and also real application

This paper is intended for preserving privacy [3] in big data healthcare application. To protect information from malicious users, Triple DES algorithm is proposed. Hereby, data will be securely stored in big data storage. The benefit of Triple DES is following: high reliability, longer key-length, and protect data and users from numerous attacks. Anonymization is applied to hide sender and receiver personal identities (name, age, mobile number, symptom of patients). On the other hand, data sharing is executed using three steps such as registration, login and authentication. Herein user name, password, secure code are hashed by *SHA-256*. Authentication the hash value is matched.

Problem

- The process of TripleDES is very slow and small block size (64 bits). However a larger block size is desirable to obtain high efficiency and security. Encryption and Decryption time is high

- User privacy may be leak in big data sharing since insufficient parameters are considered for authentication.

Proposed Solutions

- User privacy is considered in data sharing since we consider User ID, Password, Current Timestamp and Biometric for authentication. User ID and Password are cross product and then given for registration. Here we used Blake-3 algorithm, which is secure than SHA2-256

In this paper different security attacks include [4] replay, password guessing, stolen verifier, privileged-insider, denial-of-service, chosen plaintext, server-side compromisation attack, man-in-the-middle attack and so on. The proposed scheme authentications using advanced encryption standard (AES) and Elliptic Curve Cryptography (ECC).

Problem

- Hybrid AES and ECC increases encryption and decryption time for large size file
- Hadoop cluster size is large and traffic may be huge and hence AES and ECC are insufficient.

Proposed Solutions

- Cloud enabled Hadoop environment support for large traffic from users and data owners with the support of MapReduce framework

This paper proposes an integrated methodology [5] to classify and secure big data before executing data mobility, duplication and analysis. The necessity of securing big data mobility is determined by classifying the data according to the risk impact level of their contents into two categories; confidential and public. Based on the classification category, the impact of data security is studied and substantiated on the confidential data in the scope of Hadoop Distributed File System. It is revealed that the proposed approach can significantly improve the cloud systems data mobility.

Problems

- Big data classification becomes very complex since arrival rate of data storage request will be higher since after classification, deduplication is implemented

Proposed Solutions

- After determine the data deduplication, classification is implemented and then stored into HDFS, which optimizes storage of Hadoop

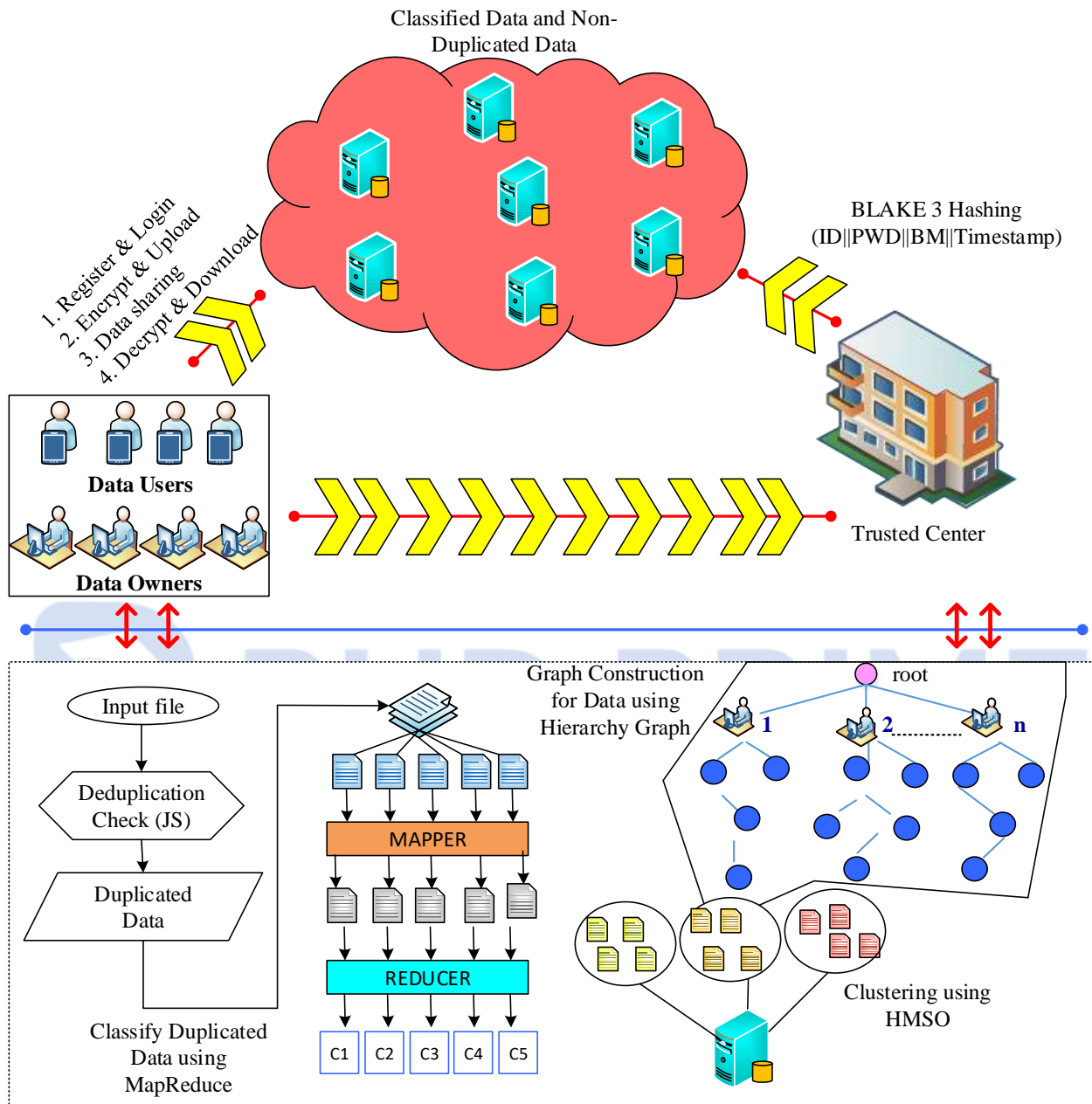
III. RESEARCH CONTRIBUTIONS

To resolve the problems of the existing approaches, in this paper we designed novel system architecture for secure data deduplication and classification by Intelligent Solutions. Our proposed work comprised of four entities including Trusted Center (TC), Data Owner (DO), Data User (DU) and Cloud Server (CS). There are three big data strategies we have proposed in this architecture:

- (i). Authentication (users and owners)
- (ii). Big Data Secure Storage
- (iii). Big Data Management and Retrieval

SYSTEM ARCHITECTURE





(i). Authentication (users and owners)

Initially, DO register their identities to TC before outsourcing data to CS. For that DO provide information of $user_{ID}$, $password$, $current\ timestamp$, & $biometric\ (eye\ vein)$ to TC. User ID and Password are cross product and then given to TC for registration. After registration, TC

generates hash value for DO given information using **BLAKE-3 hashing algorithm**. After successful authentication to the TC, DO requests private key for data encryption. Similar to data owners, data users are authenticated to the system.

(ii). **Big Data Secure Storage**

TC generates private key based on the data level requested by the DO. However, access control must be given for sensitive information. Hence, number of access of data is monitored in the CS to avoid any security breaches. TC generates keys using AES Encryption algorithm. AES-128 bits, and 256 bits of key size are used for data. Encryption of large size of data which need to outsourcing to the cloud is actually time consuming process, which takes large amount of time for encryption and decryption. It takes very less amount of time for encryption.

(iii). **Big Data Management and Retrieval**

Clustering, Indexing are plays significant role in big data storage systems. We have included these three mechanisms to further improve the storage systems. Firstly we detect duplicate before data classification. To mitigate the issues of previous works, in this work we proposed **Jaccard Similarity** for duplicate data detection. After that only we perform data classification based on the aforementioned procedure. Before storing data to the CS, it will be clustered to reduce the storage space and also searching time for both users and owners. Access control policy for DU is maintained in the CS and it updates whenever cipher text is modified by the DO. Clustering is implemented using unsupervised learning algorithm called **Human Mental search optimization**. CS is comprised of "n" number of domain servers DS_1, DS_2, \dots, DS_n . The number of clustered partitions of data is put into domain server. Each DS must maintained tree for available data partitions, which is constructed using **Hierarchy Graph**, which requires less individual searching time and proper insertions and deletion of data. It is also outperforms than the B-Tree and B+ Tree. Our proposed system solves several insider attacks over cloud enabled big data environment. The process of encryption is follows:

- (i). For input file f , DO partitioned into N number of chunks with fixed size

- (ii). These number of chunks are sent to MAPREDUCE processing framework for encryption
- (iii). Plain text chunks are sent to MAPPER for encryption which run parallel way using J48
- (iv). Classified are sent to REDUCER. Herein classified chunks are combined into single file
- (v). The single classified file is outsourced to the CS

Performance Evaluation

The performance evaluation of the proposed approach is evaluated for the following metrics:

- Throughput
 - Clustering speed (MB/s)
 - Sharing speed (MB/s)
- Encryption Time (s)
- Decryption Time (s)
- Information Loss (%)

IV. PREVIOUS WORKS & LIMITATIONS

Paper 1

Title: Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges

Concept- Big Data can be defined as the large size of datasets that are complex or difficult to process using traditional data processing application so it is created a big threat to privacy of individual. Big data analytics platforms can process data like personal information of individuals which need to be taken care when deriving some useful results for research. Over the last few decades, various privacy preserving techniques have been proposed including anonymization which needs having dataset divided in the set of attributes namely sensitive attributes, quasi identifiers and non-sensitive attributes. With structured data format, it may possible to have such a distribution but in case of unstructured data formats, it is very hard to identify the attributes.

This paper reviews the issues and challenges while privacy preserving unstructured big data analytics.

Paper 2

Title: Privacy Preserving Big Data Mining: Association Rule Hiding using Fuzzy Logic Approach

Concept – This article presented association rule mining hiding techniques to avoid the risk of sensitive knowledge leakage. For data anonymization, authors have considered association rule mining hiding technique as a sensitive association rule is applied to data with an appropriate membership degree. The main reason of association rule hiding technique is used to hide sensitive rules, without any side effect on non-sensitive rules. Along with this fuzzy logic approach is considered as a mandatory part in big data association rule hiding. Rules which are having confidence value near defined threshold are not considered as non-sensitive rules as rules with low confidence value. A rule with confidence value near the defined confidence threshold is high as a sensitive.

Limitations

- Fuzzy Logic Approach does not perform well since it leads high computation overhead and association rule hiding technique increases the copy of the data because data deduplication is necessitate for this application

Paper 3

Title: Protection of Big Data Policy

Concept- This article discussed a number of privacy-preserving mechanisms for privacy protection at different stages such as data generation, storage and data processing, User's privacy is breached due to the following circumstances: User's sensitive information are processed and stored in a location which is not properly secured, user's personal information is collected and used to add value to business. For e.g. individual's shopping habits can reveal a huge amount of

personal information and when external datasets combined with personal information which leads to the inference of new facts of the users.

Paper 4

Title – From Individual to Group Privacy in Big Data Analytics

Concept - This paper is concerned with the group privacy to balance individual privacy when assessing the ethical acceptability of analytics platforms. Due to the advances in big data analytics require new protection mechanisms for the individual privacy. The group's identity however reducible to the identities of individual members, which are defined by a set of identity tokens than those identifiers can be define the group.

Paper 5

Title – Privacy by Design in Big Data an Overview of Privacy Enhancing Technologies in the Era of Big Data Analytics

Concept - This report presented challenges and opportunities for big data technology in terms of privacy and individual's protection. This has given opportunities to very serious privacy concerns, particularly relating to wide scale electronic surveillance, disclosure and profiling of private data. The main privacy enhancing technologies were described in this paper and the special focus is on encryption, transparency, privacy by security, access control and big data anonymization mechanisms.

Paper 6

Title – Privacy-Preserving Cipher text Multi-Sharing Control for Big Data Storage

Concept - This paper presented a privacy preserving cipher text multi-sharing technique to obtain the privacy preserving properties like conditional sharing, multiple receiver-update and anonymity. Proxy Re-Encryption (PRE) is presented to solve the problem of data sharing. It consists of semi-trusted part which is called proxy. It is used transform a cipher text for a user to another user without leakage of message or decryption keys.

Limitations

- Identity based proxy Re-encryption process takes high execution time

Paper 7

Title – Big Data: Big Challenges to Privacy and Data Protection

Concept –

This paper presented the issue of privacy and data protection in big data. The EU (European Union) data protection working party illustrates about big data as the high growth in the availability and automated use of information, which protects the individual privacy. Today people from various organization, institutions and government corporations expect strong privacy protections. Although data protection and privacy is still alive and well, due to the following features: availability of data at a high scale collected on online and other devices, the use of high storage capacity, and require speed.

Paper 8

Title – Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection

Concept –

This paper addresses the potential risks and challenges related to the change of paradigm in social investigation. For that authors have created new layer which refers by the collective dimension of data protection, which protects groups of persons from the harmful threats. In this article, the assessment of the ethical and social impact of analytics to discover how to protect the collective information of groups or individuals discussed.

Paper 9

Title – Data Protection in the Context of Digital Financial Services and Big Data

Concept –

This article discussed about big data applications, especially in digital financial services. Due to the exponential growth devices such as tablets, smart phones, and PCs, the amount of data produced worldwide increasingly. Most commonly, data records are generated by digital financial sectors allows several potential risks like credit risks and insurance.

Paper 10

Title – Big Data Analysis-based Security Situational Awareness for Smart Grid

Concept –

In this paper, we proposed security situational awareness based big data analysis is demonstrated for smart grid applications. Game theory, reinforcement learning and fuzzy cluster based analytic model is combined together to propose for security analysis in smart grid. Herein, real security factors are fed into neural network as input parameters in security situational awareness model. In game theory approach, legitimate users and insider attackers are players in this game.

Limitations

- High complexity due to the involvement of deep learning and game theory approaches.

Paper 11

Title – Enhancement of data confidentiality and secure data transaction in cloud storage environment

Concept –

With the nature of centralized environment, malicious users can alter the data without the permission from the data owner. In this paper big data security is given for data owners based on Cyclic Shift Transportation Algorithm. On the other hand, hash based timestamp is proposed, which is used to stop/prevent real-time attacks. The process of involving in data owner side is follows: the given input file is partitioned into $N \times N$ matrix, then implemented shifting operations, hash code is determined for encrypted file and finally stored in cloud server.

Limitations

- The proposed approach does not consider insider attackers, which need to be considered to recover the data.

Paper 12

Title – TPTVer: A trusted third party based trusted verifier for multi-layered outsourced big data system in cloud environment

Concept –

Third party auditor plays a very important role in cloud enabled big data environment. Hereby, data owner does not audit and manage data in storage systems, but it is not an easy task for data owner to protect their trustworthiness in cloud environment. There are two policy methods introduced in which trusted data computation environment is considered in chain of trust to MapReduce application and other policy is MapReduce application is introduced for behavior measurement.

Limitations

- High complexity due to the two policy methods

Paper 13

Title – Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing

Concept –

Security Analytics is conducted over cloud of virtualized infrastructure, which is stored in the HDFS (Hadoop Distributed File System). In this paper, two step machine learning model is proposed includes logistic regression (to compute the conditional probabilities of attacks through attributes) and belief propagation (to compute the belief propagation is used to compute the belief in existence of an attack)

Limitations

- This paper does not solve some of the issues of Big Data 3V's (volume, velocity and veracity).

Paper 14

Title – A System Architecture for the Detection of Insider Attacks in Big Data Systems

Concept –

In this paper, two-step attack detection algorithm is implemented and secure communication protocol is developed to monitor the execution process of the system. The first step involves the construction control of the system for each process. In second step, instructions are matching with replica nodes.

Limitations

- In secure data communication protocol, data nodes generate random keys, which leak privacy of user and their data

Paper 15

Title – Secure Authentication in Cloud Big Data with Hierarchical Attribute Authorization Structure

Concept –

In this paper, a secure authentication protocol is proposed using tree-based signature in hierarchical attribute authorization structure. The proposed protocol can also be used in multiple-level structure for user authentication. This paper resists against forgery attack and replay attack and also protects the property of privacy preservation.

Limitations

- Hierarchical attribute authorized structure cause more time complexity and storage issues.

BIBLIOGRAPHY

- Brijesh B.Mehta., Udai Pratap Rao., (2016). Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges, Procedia Computer Science, Vol. 78, PP. 120-124
- Golnar Assadat Afzali., Shahir Mohammadi., (2017). Privacy Preserving Big Data Mining: Association Rule Hiding using Fuzzy Logic Approach, IET Information Security, PP. 1-10
- Abid Mehmood., Iynkaran Natgunanathan., Yong Xiang., Guang Hua., Song Guo., (2016). Protection of Big Data Privacy, IEEE Access, Special Section on Theoretical Foundations for Big Data Applications: Challenges and Opportunities, PP. 1-14
- Brent Mittelstadt., (2017). From Individual to Group Privacy in Big Data Analytics, Springer Philos.Technology, PP. 1-20
- Enisa Europa (2015). Privacy by Design in Big data An Overview of Privacy Enhancing Technologies in the Era of Big Data Analytics, European Union Agency for Network and Information Security
- Kaitai Liang., Willy Susilo., Joseph K.Liu., (2015). Privacy-Preserving Cipher Text Multi-Sharing Control for Big Data Storage, IEEE Transactions on Information Forensics and Security, PP. 1-11
- Abu Bakar Munir., Siti Hajar Mohd Yasin., Firdaus Muhammad Sukki., (2015). Big Data: Big data Challenges to Privacy and Data Protection, International Journal of Computer and Information Engineering, Vol. 9, No.1, PP. 1-9
- Alessandro Mantelero., (2016). Personal Data for Decisional Purposes in the Age of Analytics: From an Individual to a Collective Dimension of Data Protection, Computer Law & Security Review, Vol. 32, PP. 238-255
- Data Protection in the Context of Digital Financial Services and Big Data, Deutsche Gesellschaft for Internationale Zusammenarbeit (GIZ) GmbH, 2016

- Jun Wu, Kaoru Ota, Mianxiong Dong, Jianhua Li, Hongkai Wang, Big Data Analysis-based Security Situational Awareness for Smart Grid, IEEE Transactions on Big Data, Vol. 4, Issue.3, PP. 408-417, 2018
- Neela, K. L., & Kavitha, V. (2017). Enhancement of data confidentiality and secure data transaction in cloud storage environment. Cluster Computing, 21(1), 115–124. doi:10.1007/s10586-017-0959-4
- Zhan, J., Fan, X., Cai, L., Gao, Y., & Zhuang, J. (2018). TPTVer: A trusted third party based trusted verifier for multi-layered outsourced big data system in cloud environment. China Communications, 15(2), 122–137. doi:10.1109/cc.2018.8300277
- Win, T. Y., Tianfield, H., & Mair, Q. (2018). Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing. IEEE Transactions on Big Data, 4(1), 11–25.
- Aditham, S., & Ranganathan, N. (2017). A System Architecture for the Detection of Insider Attacks in Big Data Systems. IEEE Transactions on Dependable and Secure Computing, 1–1.
- Shen, J., Liu, D., Liu, Q., Sun, X., & Zhang, Y. (2017). Secure Authentication in Cloud Big Data with Hierarchical Attribute Authorization Structure. IEEE Transactions on Big Data, 1–1.
- Hu, C., Li, W., Cheng, X., Yu, J., Wang, S., & Bie, R. (2018). A Secure and Verifiable Access Control Scheme for Big Data Storage in Clouds, IEEE Transactions on Big Data, 4(3), 341–355. doi:10.1109/tbdata.2016.2621106
- Nafis, M. T., & Biswas, R. (2019). A secure technique for unstructured big data using clustering method, International Journal of Information Technology, Springer, doi:10.1007/s41870-019-00278-x
- Ramya Devi, R., & Vijaya Chamundeeswari, V. (2018). Triple DES: Privacy Preserving in Big Data Healthcare. International Journal of Parallel Programming. doi:10.1007/s10766-018-0592-8

- Chattaraj, D., Sarma, M., Das, A. K., Kumar, N., Rodrigues, J. J. P. C., & Park, Y. (2018). HEAP: An Efficient and Fault-tolerant Authentication and Key Exchange Protocol for Hadoop-assisted Big Data Platform. IEEE Access, 1–1
- Hababeh, I., Gharaibeh, A., Nofal, S., & Khalil, I. (2018). An Integrated Methodology for Big Data Classification and Security for Improving Cloud Systems Data Mobility. IEEE Access, 1–1

