Sensitive Data Classification & Selective Encryption Technique for Big Data

Basavaraju M M	Inchara M	KumudaShreeh H S	Dr P.C Srikanth
MTech PG Student	Mtech PG Student	Mtech PG Student	Professor and Former Head
MCE Hassan	MCE Hassan	MCE Hassan	MCE Hassan
_Dept of ECE	Dept Of ECE	Dept Of ECE	Dept Of ECE
Hassan India_	Hassan India	Hassan India	Hassan India

Abstract: Security has turned into a significant issue when the uses of huge information are drastically developing in cloud processing. The advantages of the execution for these rising innovations have enhanced or changed administration models and enhance application exhibitions in different points of view. In any case, the surprisingly developing volume of information sizes has too brought about numerous difficulties day by day. The execution time of the information encryption is one of the difficult issues amid the information preparing and transmissions. Numerous present applications surrender information encryptions so as to achieve a supportive execution level companioning with protection concerns. In this paper, we focus on protection and propose a novel information encryption approach, which is called ***Sensitive Data Classification & Selective Encryption (SCSE)***. Our proposed approach intends to specifically scramble information also, utilize protection characterization techniques under planning requirements. This approach is intended to boost the security assurance scope by utilizing a specific encryption system inside the required execution time necessities. The execution of PCSDE has been assessed in our analyses, which gives the verification of the security upgrade.

Index Terms-Cloud Processing, Big Data SCSE

I. INTRODUCTION

Presenting portable distributed computing strategies has engaged various applications throughout individuals' life as of late. Including people in the distributed computing and remote association circles turns into a shift for data recovery getting from watching people's practices and interactivities over different interpersonal organizations and portable applications. In addition, as a rising innovation, distributed computing has spread into incalculable fields with the goal that numerous new administration arrangements are acquainted with general society, for example, portable parallel processing and circulated versatile information stockpiling. Entrances of huge information strategies have additionally advanced the channels of picking up data from the expansive volume of versatile applications' information crosswise over different stages, spaces, and frameworks. Being one of specialized standards has empowered enormous information to be broadly connected in numerous mechanical areas and also investigated in late inquires about.

Regardless of numerous advantages of utilizing portable distributed computing, there are incredible worries in securing information proprietors' protection amid the interchanges on interpersonal organizations or versatile applications. One of the security concerns is caused by decoded information transmissions because of the huge volume of information. Considering a satisfactory execution level, numerous applications forsake utilizing figure messages in versatile cloud information transmissions.

This marvel can bring about protection spillage issues since plain messages are unchallenging for enemies to catch data in an assortment of routes, for example, sticking, observing, and parodying. This protection issue is urgent on the grounds that it countenances to a logical inconsistency between the security levels and execution that is typically appended to timing limitations.

II. Related Works

Cloud is a between connective stage for PC clients, which underpins data sharing among different gatherings crosswise over particular framework. Remote correspondences between clients for the most part bring individual data through an assortment of channels, including long range interpersonal communication locales and foundation. Securing information administration, stockpiling, and transmissions are three essential angles that have been investigated in the earlier inquires about.

To start with, looks into tending to the assaults in informal communities have been focused by numerous researchers. Zhang et al. proposed an approach named SCLPV for cloud-based Cyber Physical Social Systems (CPSS) to keep away from malignant evaluators. This approach simultaneously provisioned certificate less open check and in addition opposition against malevolent examiners to verify the trustworthiness of outsourced information in CPSS. Wang et al. concentrated on building up an approach offering a safe cloud framework that could bolster security protecting open inspecting. This examination had investigated the technique for characterizing foes from the information stockpiling side. In any case, the dangers in the information transmissions were not tended to with the goal that this sort of arrangement may bring about security spillages before the inspecting activities.

Besides, security concerns can be caused by different measurements in versatile mists. Dishonest information is the primary part of making security spillages that can be not really seen by clients or specialist organizations because of two principle reasons. The primary reason is that it is hard to distinguish the gathered information on account of the low dependable. The other one is that enemies don't appropriate any distinguishing proof data to such an extent that it is difficult to create risk cautions. Next, the information logical procedure is viewed as a potential answer for distinguish dependable information while the information measure turns out to be vast, which has been tended to by the earlier research. By and by, these investigates for the most part focus on danger discoveries by utilizing an assortment of examination strategies. Our examination proposes an approach that intends to build the rate of encoded information while thinking about the estimation of the information encryptions.

Be that as it may, there are assortments of vulnerabilities despite the fact that numerous entrance control models have been created. Information transmissions in remote systems make an expansive number of chances for assailants to interrupt the interchanges and take information. Protection can be debilitated even a few information portions are caught by enemies in light of the propelled information mining systems. In this paper, we build up a novel approach that specifically encode information keeping in mind the end goal to ensure protection notwithstanding amid the information transmission process. The selective data encryptions depend on the return value of the sensitive data weights and data attributes in order to reduce the chance of sensitive data leakage when hackers apply data hacking techniques.

III. Problem Definition

We describe the main problem in this system is Large Data Upload under Timing Constraints (LDuTC) problem.

Definition: Large Data Under Timing Constraints (LDuTC)

Inputs: Large data types fDig, the number of data for each data file type NDi , execution time when encrypting data for each single data TeD i , execution time without encryptions for each single data TnDi , the privacy weight value for each data type WDi .

Outputs: a technique classify which file will be encrypted. The proposed problem is finding out the approach that can gain the maximum total privacy weight value under a given timing constraint.

As illustrated in Definition, the main inputs include five variables. First, input data include a group of files that are classified into different types, represented as a set fDig. The number of data file in each type Di is represented as NDi This system has two different uploading process, one process is Upload with Encryptions (UwE) and another is Upload with Non-Encryption (UwNE). The execution time of each data file Di in UwE mode is TeDi . Similarly, the execution time of each data package Di in UwNE mode is TnDi . Furthermore, we introduce a parameter, Sensitive Weight Value (PWV), for each data file in order to calculate to classify whether the file is sensitive or not before encrypting data, represented as WDi .

IV. PROPOSED SYSTEM

Sensitive Data Classification & Selective Encryption (SCSE)

Based on the problem definition, we present out SCSE system in this section. The main focus of this system is to solve the problem in Large Data Under Timing Constraints (LDuTC).

In this system there are two actors one is Admin and another is User, Admin has to provide the sensitive data training set to the system. Data user can able to register and login with the registration credentials. Once data user logged in he can able to upload and download the file as he wish. Fig.1 shows this process clearly.

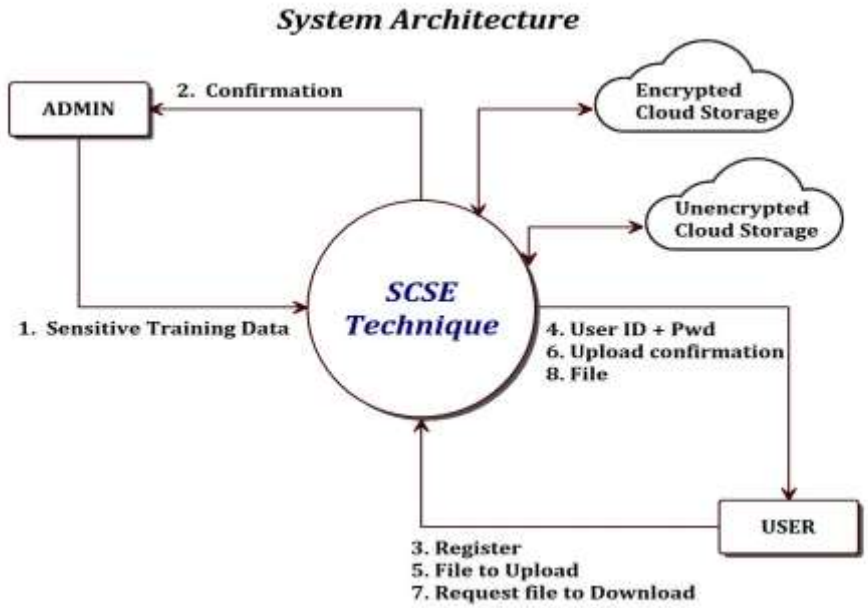

Fig 1: SCSE System Architecture

When the data user upload the file, The following three process are take place

1. Keyword Extraction
2. Sorting Keywords
3. Decision Making

While user uploading a file, unnecessary characters and words are removed then using Term Frequency (TF) algorithm the keywords are extracted then keyword weights are calculated from the uploaded file.

Extracted keywords then enter into sorting process where keywords are compared with sensitive keywords which are provided by Admin, the assumption is that admin has to upload all sensitive data into the system well in advance, the output of this process gives filtered sensitive keywords in uploaded file.

The third process is decision making which decides whether to encrypt the file is not based on the filtered sensitive keyword accumulative weight. The accumulated weight of sensitive keywords are check with predefined threshold, if it cross the threshold value then encryption will take place else the file uploaded without encryption. Fig.2 shows the process in detail.

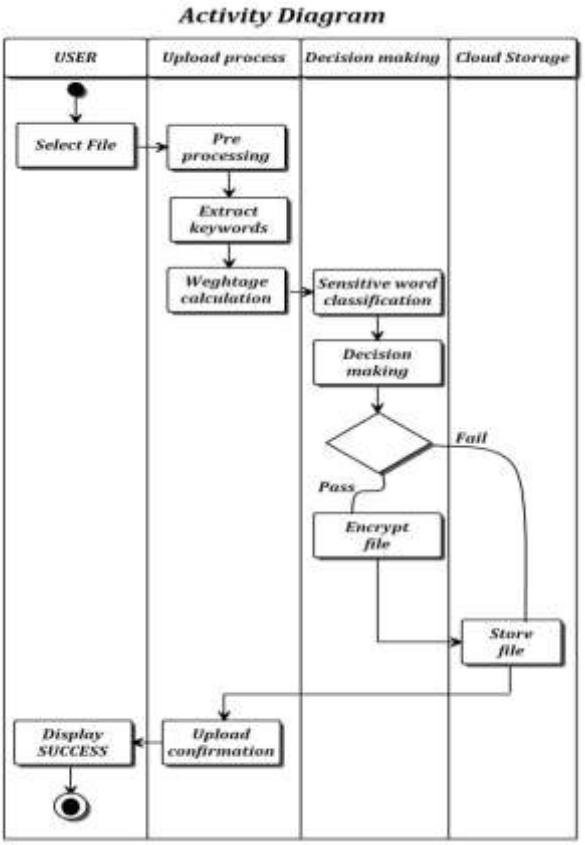

Fig 2: Activity Diagram for upload process

V. CONCLUSION

This paper concentrated on the protection issues of enormous information and thought about the down to earth executions in distributed computing. The proposed approach, SCSE, was intended to expand the productivity of security insurances. Main algorithm supporting SCSE model was decision making algorithm that was developed to dynamically alternative data files for encryptions under different timing constraints. The exploratory assessments demonstrated the proposed approach had a versatile and unrivaled execution.

REFERENCES

[1] S. Yu, W. Zhou, S. Guo, and M. Guo. A feasible IP traceback framework through dynamic deterministic packet marking. IEEE Transactions on Computers, 65(5):1418–1427, 2016.
[2] S. Yu, G. Gu, A. Barnawi, S. Guo, and I. Stojmenovic. Malware propagation in large-scale networks. IEEE Transactions on Knowledge and Data Engineering, 27(1):170–179, 2015.
[3] S. Liu, Q. Qu, L. Chen, and L. Ni. SMC: A practical schema for privacy-preserved data sharing over distributed data streams. IEEE Transactions on Big Data, 1(2):68–81, 2015.
[4] S. Rho, A. Vasilakos, and W. Chen. Cyber physical systems technologies and applications. Future Generation Computer Systems, 56:436–437, 2016.
[5] L. Wu, K. Wu, A. Sim, M. Churchill, J. Choi, A. Stathopoulos, C. Chang, and S. Klasky. Towards real-time detection and tracking of spatio-temporal features: Blob-filaments in fusion plasma. IEEE Transactions on Big Data, 2(3), 2016.
[6] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar. Dependable demand response management in the smart grid: A stackelberg game approach. IEEE Transactions on Smart Grid, 4(1):120–132, 2013.
[7] M. Qiu, M. Zhong, J. Li, K. Gai, and Z. Zong. Phase-change memory optimization for green cloud with genetic algorithm. IEEE Transactions on Computers, 64(12):3528–3540, 2015.
[8] H. Liu, H. Ning, Y. Zhang, Q. Xiong, and L. Yang. Role-dependent privacy preservation for secure V2G networks in the smart grid. IEEE Transactions on Information Forensics and Security, 9(2):208–220, 2014.
[9] F. Tao, Y. Cheng, D. Xu, L. Zhang, and B. Li. CCIoT-CMfg: cloud computing and internet of things-based cloud manufacturing service system. IEEE Transactions on Industrial Informatics, 10(2):1435–1442, 2014.
[10] G. Wu, H. Zhang, M. Qiu, Z. Ming, J. Li, and X. Qin. A decentralized approach for mining event correlations in distributed system monitoring. Journal of parallel and Distributed Computing, 73(3):330–340, 2013.

[11] S. Yu,W. Zhou, R. Doss, andW. Jia. Traceback of DDoS attacks using entropy variations. IEEE Transactions on Parallel and Distributed Systems, 22(3):412–425, 2011.
[12] Y. Li, W. Dai, Z. Ming, and M. Qiu. Privacy protection for preventing data over-collection in smart city. IEEE Transactions on Computers, 65:1339–1350, 2015.
[13] S. Yu,W. Zhou,W. Jia, S. Guo, Y. Xiang, and F. Tang. Discriminating DDoS attacks from flash crowds using flow correlation coefficient. IEEE Transactions on Parallel and Distributed Systems, 23(6):1073–1080, 2012.
[14] Y. Yu, M. Au, G. Ateniese, X. Huang, W. Susilo, Y. Dai, and G. Min. Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. IEEE Transactions on Information Forensics and Security, PP(99):1, 2016.
[15] L. Weng, L. Amsaleg, A. Morton, and S. Marchand-Maillet. A privacy-preserving framework for large-scale content-based information retrieval. IEEE Transactions on Information Forensics and Security, 10(1):152–167, 2015.
[16] K. Gai, M. Qiu, H. Zhao, and J. Xiong. Privacy-aware adaptive data encryption strategy of big data in cloud computing. In The 2nd IEEE International Conference of Scalable and Smart Cloud (SSC 2016), pages 273–278, Beijing, China, 2016. IEEE.
[17] Y. Zhang, C. Xu, S. Yu, H. Li, and X. Zhang. SCLPV: Secure certificateless public verification for cloud-based cyber-physical social systems against malicious auditors. IEEE Transactions on Computational Social Systems, 2(4):159–170, 2015.
[18] C. Wang, S. Chow, Q. Wang, K. Ren, and W. Lou. Privacy-preserving public auditing for secure cloud storage. IEEE Transactions on Computers, 62(2):362–375, 2013.
[19] K. Gai, L. Qiu, M. Chen, H. Zhao, and M. Qiu. SA-EAST: security aware efficient data transmission for ITS in mobile heterogeneous cloud computing. ACM Transactions on Embedded Computing Systems, 16(2):60, 2017.
[20] C. Lai, M. Chen, M. Qiu, A. Vasilakos, and J. Park. A RF4CEbased remote controller with interactive graphical user interface applied to home automation system. ACM Transactions on Embedded Computing Systems, 12(2):30, 2013.
[21] S. Backhaus, R. Bent, J. Bono, R. Lee, B. Tracey, D. Wolpert, D. Xie, and Y. Yildiz. Cyber-physical security: A game theory model of humans interacting over control systems. IEEE Transactions on Smart Grid, 4(4):2320–2327, 2013.
[22] K. Gai, M. Qiu, H. Zhao, and W. Dai. Privacy-preserving adaptive multi-channel communications under timing constraints. In The IEEE International Conference on Smart Cloud 2016, page 1, New York, USA, 2016. IEEE.
[23] L. Tang, X. Yu, Q. Gu, J. Han, G. Jiang, A. Leung, and T. Porta. A framework of mining trajectories from untrustworthy data in cyberphysical system. ACM Transactions on Knowledge Discovery from Data, 9(3):16, 2015.
[24] F. Schuster, M. Costa, C. Fournet, C. Gkantsidis, M. Peinado, G. Mainar-Ruiz, and M. Russinovich. VC3: Trustworthy data analytics in the cloud using SGX. In IEEE Symposium on Security and Privacy, pages 38–54, San Jose, CA, USA, 2015. IEEE.
[25] M. Maffei, G. Malavolta, M. Reinert, and D. Schroder. Privacy and access control for outsourced personal records. In IEEE Symposium on Security and Privacy, pages 341–358, San Jose, CA, USA, 2015. IEEE.
[26] Y. Li, K. Gai, Z. Ming, H. Zhao, and M. Qiu. Intercrossed access control for secure financial services on multimedia big data in cloud systems. ACM Transactions on Multimedia Computing Communications and Applications, 12(4s):67, 2016.
[27] C. Mulliner, W. Robertson, and E. Kirda. Hidden GEMs: Automated discovery of access control vulnerabilities in graphical user interfaces. In IEEE Symposium on Security and Privacy, pages 149–162, San Jose, CA, USA, 2014. IEEE.
[28] S. Sen, S. Guha, A. Datta, S. Rajamani, J. Tsai, and J. Wing. Bootstrapping privacy compliance in big data systems. In IEEE Symposium on Security and Privacy, pages 327–342, San Jose, CA, USA, 2014. IEEE.
[29] J. Vilk, D. Molnar, B. Livshits, E. Ofek, C. Rossbach, A. Moshchuk, H. Wang, and R. Gal. SurroundWeb: Mitigating privacy concerns in a 3D web browser. In IEEE Symposium on Security and Privacy, pages 431–446, San Jose, CA, USA, 2015. IEEE.
[30] L. Zhu, Z. Hu, J. Heidemann, D. Wessels, A. Mankin, and N. Somaiya. Connection-oriented DNS to improve privacy and security. In IEEE Symposium on Security and Privacy, pages 171–186, San Jose, CA, USA, 2015. IEEE.
[31] M. Baran, P. Carpenter, L. Borbye, D. Lubkeman, M. Ligett, and D. Covington. A new professional science master program for electric power systems engineering. IEEE Transactions on Power Systems,
29(4):1903–1910, 2014. [32] K. Gai and A. Steenkamp. A feasibility study of Platform-as-a- Service using cloud computing for a global service organization. Journal of Information System Applied Research, 7:28–42, 2014.
[33] K. Li, W. Zhang, C. Yang, and N. Yu. Security analysis on one-tomany order preserving encryption-based cloud data search. IEEE Transactions on Information Forensics and Security, 10(9):1918– 1926, 2015.
[34] C. Cicconetti, L. Lenzini, A. Lodi, S. Martello, E. Mingozzi, and M. Monaci. Efficient two-dimensional data allocation in IEEE 802.16 OFDMA. IEEE/ACM Transactions on Networking, 22(5):1645–1658, 2014.
[35] M. Vidyasagar. A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. IEEE Transactions on Automatic Control, 57(10):2464–2477, 2012.
[36] M. Qiu, Z. Chen, Z. Ming, X. Qin, and J. Niu. Energy-aware data allocation with hybrid memory for mobile cloud systems. IEEE Systems Journal, PP:1–10, 2014.
[37] P. Lacroute. Real-time volume rendering on shared memory multiprocessors using the shear-warp factorization. In Proceedings of the IEEE symposium on Parallel rendering, pages 15–22, Atlanta, GA, USA, 1995. ACM.