ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(05), May 2023, pp. 1-8

Article Title: Effective and privacy protecting cross domain deduplication in cloud

Effective and Privacy Protecting Cross Domain Deduplication in Cloud

E. Vino Dharshini[1], K. Vahitha Thangam[2]

[1]PG Student, Department of Computer Science and Engineering, Rohini College of Engineering and Technology, Kanyakumari, India, vinodharshini5@gmail.com
[2]Assistant Professor, Department of Computer Science and Engineering, Rohini College of Engineering and Technology, Kanyakumari, India, vahisudhan5@gmail.com

ABSTRACT

In the recent years, the big data world has developed a cloud with strong storage management that can verify data integrity and maintain one data duplicate. The data deduplication (DD) issue has been solved by the development of numerous cloud auditing storage techniques, but these methods are weak and unable to withstand brute force attacks. In this paper, the proposed technique explore a three-tier cross-domain architecture and suggest a fast and private huge data deduplication in cloud storage. EPCDD achieves data availability as well as privacy preservation while resisting brute-force attacks. In order to provide better privacy protections than previous systems, this method take accountability into consideration. In terms of compute, communication and storage overheads and then it show that EPCDD performs better than currently used competitive strategies. However, since users and data owners may not be confident in cloud storage providers, data will likely be encrypted before outsourcing. Since different users encrypt identical data in different ways, deduplication efforts are complicated. As a result, to examine the performance of the proposed work is utilizing the java software.

Keywords: Data deduplication, EPCDD, Data availability, Java software, cloud auditing.

1 Introduction

Several technologies have emerged in the 21st century which have fundamentally changed the way that people interact with the current digital ecology. One of the most important of these is cloud storage. The possibility of using that data across different devices and for different reasons has simply grown because peoples save their data remotely instead of physically storing it on their devices [1-2]. Data deduplication technology has been widely developed in cloud storage in the big data-driven in the world because it greatly minimize storage costs by storing only a single copy of redundant data [3]. In order to save space and improve storage performance, cloud storage desperately uses effective deduplication strategies, which implies that cloud servers only keep a new instance of any duplicate file and the service provider needs a connection to access a single file on all DO's in the same file [4-5]. The issue of security is addressed by cloud-based data services, which may result in effective security of data that is moved and exchanged among users for specific applications. Moreover, the service improve QoS and offer better privacy for private and confidential data. Based on the

efficient use of data specific to resource sharing, the cloud offers a variety of services to its users [6-7]. The deduplication technique divided into client-side as well as server-side approaches. There has been a lot of interest in edge computing with deduplication. But, there are still many security issues and potential threats in academia and industry [8].

Medical data in both image and text format is used in cloud-based e-healthcare systems that protect patient privacy. Sensor devices routinely gather data, which is then processed by mobile devices. Before sent to the cloud, this data must be encrypted because the cloud is secure but curious may seek to collect personal details. Yet, because the mobile device has limited resources, the encryption should be minimally computationally intensive and lightweight [9-10]. Because of these concerns, some customers decide against entrusting their files to cloud service providers and instead manage and keep them locally. This inherently creates problems and charges that increase as their data grows more difficult. This segment of consumers is also in danger due to the rise in malware, particularly ransom ware that is growing increasingly sophisticated [11-12]. This method is observable that it is difficult to create an effective deduplication method which achieves privacy preservation, availability and accountability while fending off brute-force attacks. Hence, it suggest an effective and privacy-preserving big data deduplication in cloud storage, referred to as EPCDD, in this study employing a three-tier cross-domain architecture. The EPCDD method successfully maintains privacy while ensuring data availability, accountability, and resistance to brute-force attacks [13-15].

The suggested solution showed that it improved data availability, accountability and privacy preservation while fending off brute-force attacks and further it showed that, in terms of compute, communication, and storage overheads, the suggested method exceeds current state-of-the-art schemes. Moreover, the duplicate search time complexity in our technique is an effective logarithmic time. Extending the suggested approach to entirely safeguard the duplicate information from disclosure, even by a hostile CSP, without compromising the capacity to perform data deduplication, becomes an area for further research. As a result, to examine the performance of the proposed work is utilizing the java software.

2 Proposed Methodology

The proposed work involves the users registering their personal information in the cloud. These details consist of the person's name, address, email address, password and gender. Each user receives a special private key or ID to access the cloud storage following provider verification. These private keys are secretly provided to each recipient's email address. Users has upload and download files from the cloud using the cloud storage provider's cloud space. Moreover, the proposed scheme fully protect the duplicate information from disclosure, even by a malicious CSP, without affecting the capability to perform data deduplication, an effective and privacy protecting data deduplication using brute force attack algorithm.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(05), May 2023, pp. 1-8

Article Title: Effective and privacy protecting cross domain deduplication in cloud

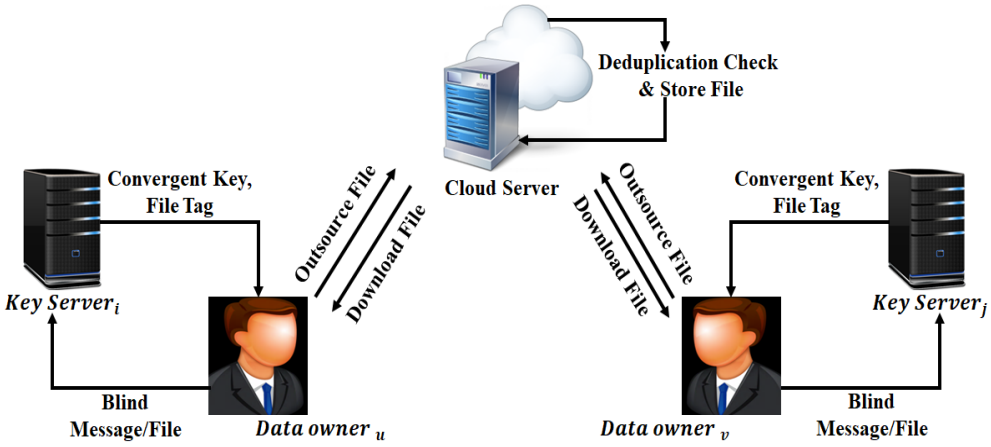

Figure 1: *Proposed block diagram*

The entire information of time period specified in the cloud is maintained by the cloud storage provider. When a private key is present, users can connect in to the cloud using their unique username and password. Each user has their own unique email address, password and private key to log into the system. Users are able to upload files to cloud storage. These files are kept encrypted on the cloud. The data in the cloud is encrypted using a cutting-edge encryption standard technique. Because a secret key has been set to the encrypted file, if there is any danger of hacking occurring during the transfer, the content cannot be viewed.

2.1 Threat Model

Theoretically, they might conspire with the CSP to gain the privacy of other clients. In reality, such collusion could lead to civil lawsuits, criminal investigations and serious consequences to the CSP's image. Additionally, the CSP is probable to work together with client B or other clients to compromise the privacy of other current clients if it works with client A to compromise the privacy of client B. If such collaboration is reported or discovered, this could have major consequences for the CSP. Therefore, assume that the CSP does not conspire with its customers as well as it not consider any other active attacks outside brute-force attacks.

2.2 Data Encryption

The client calculates the message-dependent symmetric key ski = h1 (mike (g, g) st by using secret key s and parameter e (g, g) t. Then, using the symmetric encryption process known as cypher block chaining (CBC) mode, or AESCBC, this client selects a random integer ri∈ ZN then computes the ciphertext Ci = Encski (rikmi).

2.3 Tags Generation

The client creates two tags for data mi as τ 1 i = g s·h2(mi) , τ 2 i = ski mod ω, where I is a random integer which not only ensures the security of ski but also satisfies the CSP's storage requirements. For instance, by setting || = 128 bits, ski may represent up to 2 128 bits of data and therefore it is protected from attacks if n is set to 256 bits (|ski |=256). Similar procedures

are carried out by clients in domain B to produce ciphertexts and tags, for example, encrypting mj as Cj = Encskj (rjkmj), where skj = h1 (mjke (g, g) st), and computing the accompanying tags as τ 1 j = g t·h2 (mj) ∈ G, τ 2 j = skj mod ω.

2.4 Deduplication Decision Tree (DDT) Initialization

This method develop the deduplication decision trees (DDTs) based on the well-liked binary search tree (BST) for discovering duplicate data in order to increase the effectiveness of detecting duplicated data. As far as peoples have to aware, the DDT initialization is comparable to the BST insertion, but this process starts with an empty tree. Imagine that at this point in time the CSP has received k separate message tuples from domain A, such as {(C1, τ 1 1, τ 2 1), (C2, τ 1 2, τ 2 2). . . (Ck, τ 1 k, τ 2 k)}. In order to keep k message tuples for subsequence deduplication, CSP creates a DDT-A for this domain. Moreover, it must balance the tree during the DDT construction process to guarantee that the time complexity is searching for duplicates is O (log k).

2.5 Uploading Data

Let's assume a client from domain B wants to send the data m to the CSP. This client generates two tags, 1 and 2 and sends them to the LMB. The two tags are τ 1 ∗ = g t·h2 (m∗) and τ 2 ∗ = h1 (m∗ke (g, g) st) mod ω, respectively. When receiving tags, LMB calculates the hash value for "1" (i.e., T = h3("1")) then it searches the hash table which stores hash values of the first tag for all different data from domain B. The LMB returns a "duplicate find" to the client and it not forward any messages to the CSP if the same hash value has already been logged. In any other case, LMB sends the tags (#1 and #2) to the CSP. Following receipt, CSP examines the duplication on the DDT-A. If there are duplicated data, CSP informs the LMB and transmits "duplication detect" to it. In all other cases, it "uploads data" to the LMB. LMB transmits the feedback to the client after obtaining it. The client encrypts the message "upload data" as C = Encsk (rkm), and then sends it to the CSP through the LMB. The message tuple (C, 1, and 2) is inserted into the proper node in the DDT-B by the CSP. It's essential to keep in mind that domain A and domain B both use the same upload mechanism for client data.

2.6 Privacy Analysis

Users examine how our EPCDD system will prevent the exposure of sensitive information while minimising the disclosure of redundant information. Clients are required to upload the encrypted data but also two associated tags in order to work with the CSP to process data deduplication. The symmetric encryption algorithm, AESCBC, is used to encrypt Ci, hence the security of Ci is predicated on this algorithm. Furthermore, dealing with the discrete logarithm (DL) issue and the one-way hash function, both which have been demonstrated to be computationally infeasible challenging problems, which is required if the CSP and LMA (LMB) are to extract mi from the tag τ 1 i = g s·h2 (mi) (τ 1 i = g t·h2 (mi)). As a result, mi not be obtained from τ 1 i. Because τ 2 i = ski mod ω. There is an integer k such that ski = kω + τ 2 i. CSP or LMA (LMB) only determine the solution to one equation with two unknown

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(05), May 2023, pp. 1-8

Article Title: Effective and privacy protecting cross domain deduplication in cloud

numbers by guessing at it. Users will set || = 128 bits if |ski | = 256 bits, which is adequate for avoiding a guessing attack. As a result, this paper achieve data secrecy. Moreover, to confirm that the various ciphertexts translate into the same plaintext. As only the CSP has access to the secret parameters g aq and g bq under our EPCDD scheme, it has capable to perform this verification. In other words, the duplicate information is only known to the CSP. As a result, this approach will minimise the disclosure of redundant information.

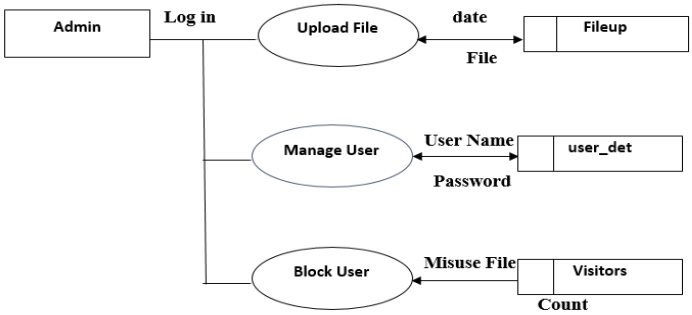

Figure 2: *Data flow diagram*

3 Results and Discussion

A Java platform-based programme is shown in the following figure. The virtual machine and Java API shield the programme from the hardware, as seen in the below figures. Native code is code which, once it has been compiled and it will execute on a particular hardware platform. The Java platform, which is a platform-independent environment, which sometimes slightly slower than native code. Instead of risking portability, smart compilers, fine-tuned interpreters and just-in-time byte code compilers will achieve performance which rivals native code.

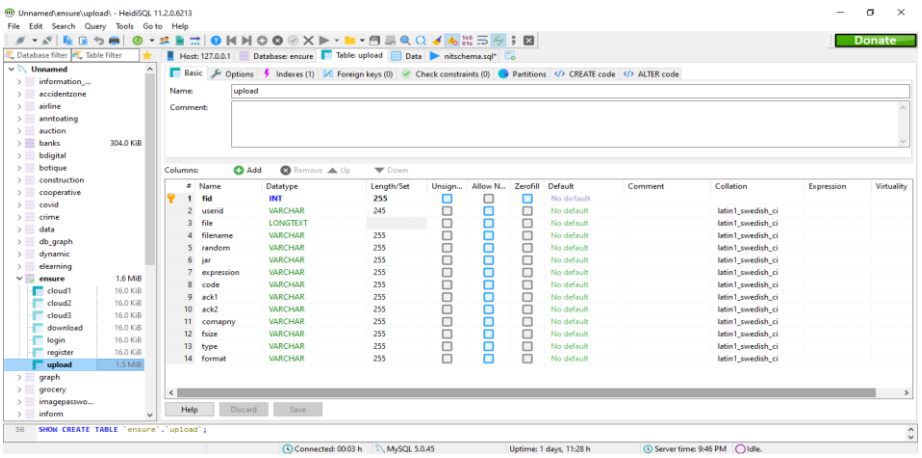

Figure 3: *Implementation Design*

Figure 3 illustrates the implementation design, from the figure it is observed that the personal details of the user profiles are ensure to upload in the file with individuals name, data types etc. to secure the personal details of the user. As represented in figure 4.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(05), May 2023, pp. 1-8

Article Title: Effective and privacy protecting cross domain deduplication in cloud

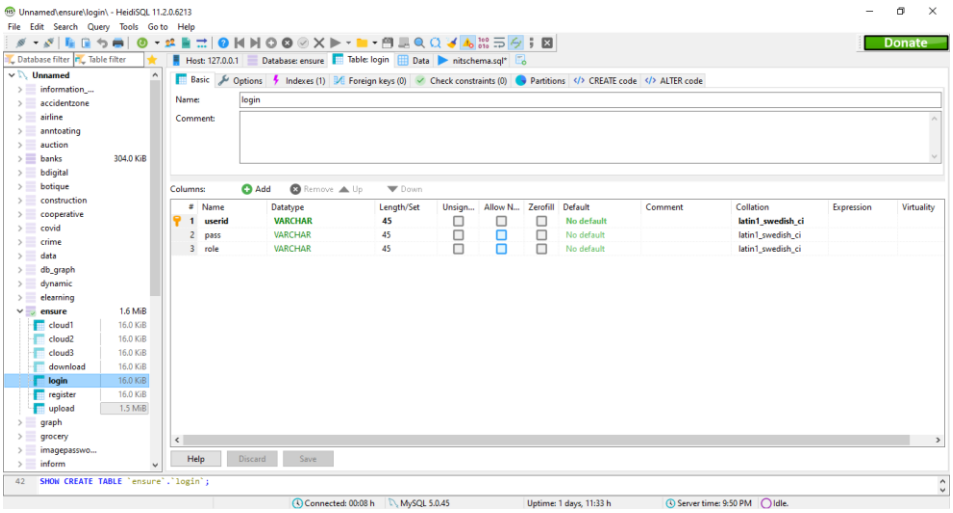

Figure 4: *Login page*

The login page is represented in figure 4, from the above figure it is noted that the personal data of the user's details are stored to ensure with login file by the individuals name, data type, length etc., which is shown by the above figure.

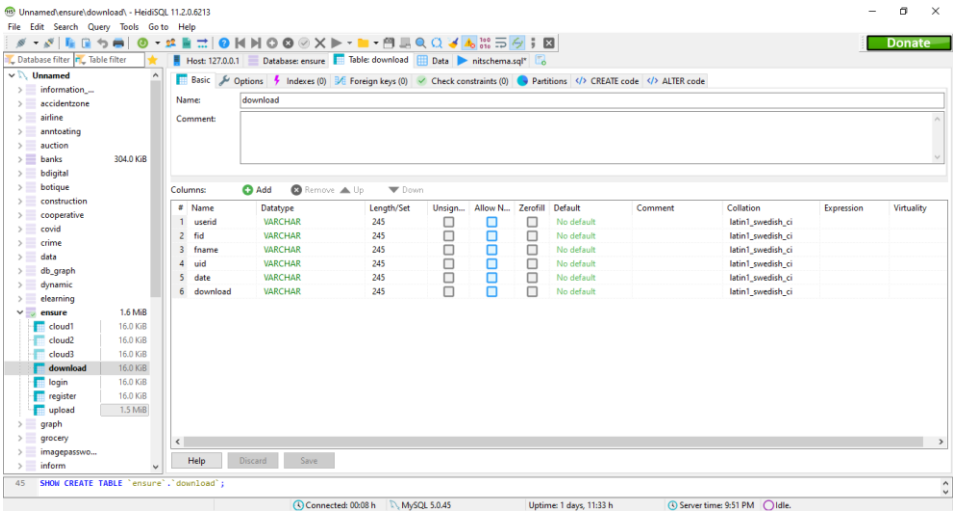

Figure 5: *Download page*

Figure 5 illustrates the download page, from the observation it is noted that the personal details of individuals data are stored in ensure with download files by the user names, data type, length etc. for secure the personal details of the users as represents in above figure 5.

The cloud server is illustrated in figure 6, from the figure it is noted that the personal details of individual users data are stored to ensure with cloud 2 file by the user name, dataset, length etc. for secure the individuals personal details as represented in above figure 6.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(05), May 2023, pp. 1-8

Article Title: Effective and privacy protecting cross domain deduplication in cloud

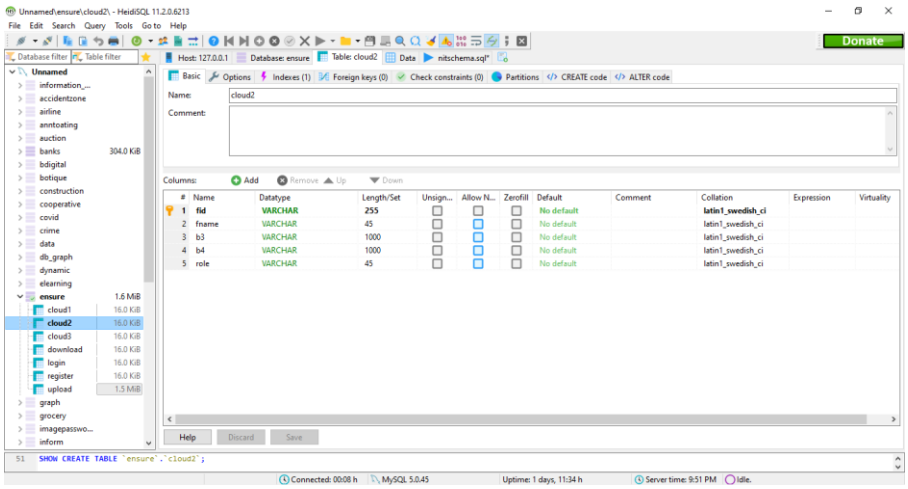

Figure 6: *Cloud server*

4 Conclusion

In this research, the suggested method present a massive data deduplication strategy for a two-level multi-domain architecture which is effective and privacy-preserving. By producing a fixed number of ciphertexts, our system secure encrypted data semantically while ensuring that users with access to the data will correctly decrypt it. The suggested a system and showed that, while resisting brute-force attacks, it improves privacy preservation, accountability and data availability. Additionally it showed that the suggested scheme performs better in terms of compute, communication and storage overheads than the current state-of-the-art approaches. Also, our scheme's duplicate search has an effective logarithmic time complexity. Further study will involve expanding the suggested system to completely shield the duplicate data from exposure, even by an evil CSP, while maintaining the capacity to execute data deduplication. Along with enhancing the temporal complexity of duplicate search, future research will also focus on expanding the scheme's resilience against a wider range of security vulnerabilities posed by outside attackers.

References

1. Gund Avinash B; Prerna N. Mahadik; Ashvini R. Thorat; Ganesh K. Yevle, Year: 2022, “Data De-Duplication Using Blockchain with Advanced Security in Cloud Computing," Available at SSRN 4289505.

2. Liu Jiasen; Xu An Wang; Zhiquan Liu; Han Wang; Xiaoyuan Yang, Year: 2020, “Privacy-Preserving Public Cloud Audit Scheme Supporting Dynamic Data for Unmanned Aerial Vehicles”, IEEE Access, Vol: 8, pp. 79428 – 79439.

3. Zhang Chen; Yinbin Miao; Qingyuan Xie; Yu Guo; Hongwei Du; Xiaohua Jia, Year: 2022, “Privacy-Preserving Deduplication of Sensor Compressed Data in Distributed Fog Computing”, IEEE Transactions on Parallel and Distributed Systems 33, no: 12, pp. 4176-4191.

4. Gnana Jeslin, J; P. Mohan Kumar, Year: 2022, "Decentralized and Privacy Sensitive Data De-Duplication Framework for Convenient Big Data Management in Cloud Backup Systems", Symmetry, Vol: 14, no: 7, pp. 1392.

5. Yang Caixia; Liang Tan; Na Shi; Bolei Xu; Yang Cao; Keping Yu, Year: 2022, "AuthPrivacyChain: A Blockchain-Based Access Control Framework With Privacy Protection in Cloud", IEEE Access, Vol: 8.

6. Bharanidharan M; Karunakaran E, Year: 2019, "An Efficient Privacy-Preserving Data De-Duplication in Cloud", International Journal of Applied Engineering Research, Vol: 8, pp. 1798-1804.

7. Li Jung-Shian; I-Hsien Liu; Chin-Jui Tsai; Zhi-Yuan Su; Chu-Fen Li; Chuan-Gang Liu, Year: 2022, "Secure Content-Based Image Retrieval in the Cloud with Key Confidentiality", IEEE Access, Vol: 08, pp. 114940 – 114952.

8. Zhou Yukun; Zhibin Yu; Liang Gu; Dan Feng, Year: 2022, "An efficient encrypted deduplication scheme with security-enhanced proof of ownership in edge computing", Bench Council Transactions on Benchmarks, Standards and Evaluations, Vol: 2, no: 2, pp. 100062.

9. Gayathri, S; S. Gowri, Year: 2022, "CUNA: A privacy preserving medical records storage in cloud environment using deep encryption", Measurement: Sensors 24, pp.100528.

10. Yong Li; Liu Hefei; Shen Xiujuan; Yuan Bin; Wang Kun, Year: 2021, "Keyword Semantic Extended Top-k Cipher text Retrieval Scheme Over Hybrid Government Cloud Environment", IEEE Access, Vol: 09, pp. 155249 – 155259.

11. Zhang Chen; Yinbin Miao; Qingyuan Xie; Yu Guo; Hongwei Du; Xiaohua Jia, Year: 2022, "Privacy-Preserving Deduplication of Sensor Compressed Data in Distributed Fog Computing", IEEE Transactions on Parallel and Distributed Systems, Vol: 33, no: 12 , pp. 4176-4191.

12. Goldstein Anat; Lior Fink; Gilad Ravid, Year: 2022, "A Cloud-Based Framework for Agricultural Data Integration: A Top-Down–Bottom-Up Approach", IEEE Access, Vol: 10, pp. 88527 – 88537.

13.He Heng; Renju Chen; Chengyu Liu; Ke Feng; Xiaohu Zhou, Year: 2021, "An Efficient Cipher text Retrieval Scheme Based on Homomorphic Encryption for Multiple Data Owners in Hybrid Cloud", IEEE Access, Vol: 09, pp. 168547 – 165887.

14. Gai Keke; Meikang Qiu; Xiaotong Sun; Hui Zhao, Year: 2016, "Smart data deduplication for telehealth systems in heterogeneous cloud computing", IEEE Access, Vol: 01, pp. 93 – 104.

15. Filip Ion-Dorinel; Andrei Vlad Postoaca; Radu-Dumitru Stochitoiu; Darius-Florentin Neatu; Catalin Negru; Florin Pop, Year: 2019, "Data capsule : representation of heterogeneous data in cloud – edge computing", IEEE Access, Vol: 07, pp. 49558 – 49567.