International Journal for Modern Trends in Science and Technology, 8(09): 167-172, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0809035
Available online at: http://www.ijmtst.com/vol8issue09.html

Spotting Suspecious Cloud File Replication or Migration

K Hruthpadma Suhasini[1] | B. N. Srinivasa Gupta [2]

[1]PG Scholar, Department of Computer Science, SVKP &Dr K S Raju Arts & Science College(A), Penugonda, W.G.Dt., A.P, India
[2]Associate Professor in Computer science, SVKP &Dr K S Raju Arts & Science College(A), Penugonda, W.G.Dt., A.P, India

To Cite this Article
K Hruthpadma Suhasini and B. N. Srinivasa Gupta. Spotting Suspecious Cloud File Replication or Migration. International Journal for Modern Trends in Science and Technology 2022, 8(09), pp. 167-172. https://doi.org/10.46501/IJMTST0809035

Article Info
Received: 22 August 2022; Accepted: 12 September 2022; Published: 17 September 2022.

ABSTRACT

In recent years, cloud storage has experienced a rapid increase in popularity. Although cloud storage has several benefits. Users are frequently unable to discern benefits, such as flexibility and convenience or have access to where their data is actually located. This restriction might impact users' faith and confidence in the storage provider, or even For data storage with precise location requirements, the cloud is inadequate. To We suggest a system named LAST-HDFS to address this problem. Includes the open source Location-Aware Storage Technique (LAST) Hadoop Distributed File System as a source (HDFS). Location-aware file allocations are required, and the LAST-HDFS system continuously checks for file transfers to look for possible illegal cloud transfers. Illegal. Here, transfers are attempts to transmit private information outside of ("legal") restrictions described by the file owner and its policies. Our underlying algorithms model file transfers among nodes as a weighted graph, and maximize the probability of storing data items of similar privacy preferences in the same region. We equip each cloud node with a socket monitor that is capable of monitoring the real-time communication among cloud nodes. Based on the real-time data transfer information captured by the socket monitors, our system calculates the probability of a given transfer to be illegal. We have implemented our proposed framework and carried out an extensive experimental evaluation in a large-scale real cloud environment to demonstrate the effectiveness and efficiency of our proposed system.

1. INTRODUCTION

The need for cloud storage has grown dramatically along with cloud computing's rising popularity. The only customers of cloud storage and cloud computing are no longer computing companies; instead, common enterprises and even end users are benefiting from the vast capabilities that cloud services may offer. While taking advantage of the flexibility and convenience offered by cloud storage, users often relinquish control over their data and are unable to find it, whether it is located locally, nationally, or even internationally.

For cloud users (such as hospitals) who store sensitive data (such as medical records) that are required by law to remain inside specific geographic boundaries and borders, a lack of geographical control may result in privacy violations. Another instance where this issue occurs is with governmental organisations that need all data to be stored in the same nation that the government is based in; this difficulty has run into problems with

cloud service providers (CSPs) covertly transferring data abroad or being acquired by foreign firms. For instance, Canadian law requires that personally identifiable information be maintained within the country. However, given that the Amazon Cloud has more than 40 zones spread out globally [1], it is exceedingly difficult to guarantee regulatory compliance. Even Hadoop, which was formerly maintained as a distributed filesystem with a limited geographic distribution, is now widely used in many areas (see Facebook Prism [2] or recent patent [3]). Theexact location of data saved in the cloud can currently be verified by a number of techniques, with a concentration on post-allocation compliance [4]–[6]. However, Recent studies have recognised the value of proactive location control for data placement that complies with adopters' location criteria [4], [7], [8], to give users more control over their data and to ensure the location where the data is stored.

In this study, we hack into Hadoop Distributed File System (HDFS), one of the most extensively used cloud data storage systems, and create LAST-HDFS, an improved HDFS system. Using the LAST HDFS, location-aware file allocations and file transfer monitoring are now possible with HDFS. LAST-HDFS specifically offers the new functions listed below: I consistently enforces a location-aware dataloading and storage by allocating data nodes in accordance with user-specified privacy policies; (ii) actively monitors and dynamically corrects potential data migration (due to balancing or data replication needs within the cluster that might violate dataplacement policies); (iii) detects possible data migration violations.

The remainder of the essay is structured as follows. Section 2 presents the use case of our suggestedsolution. Section 3briefly discuss the background concerning the Hadoop system. The suggested LAST-HDFS system is then described in Section 4 before Section 5 presents specific implementation strategies. Section 6 reports the experimental results. Finally, Section 8 ends the paper and describes potential future research possibilities.

2. LITERATURE SURVEY

While there are several published surveys on different aspects of cloud computing, to the best of our knowledge, there is not a survey devoted to review the literature on existing approaches, resulting in a process model for moving legacy applications to the cloud. Perhaps, the closest studies to this paper are those related to migrating legacies to SOA because SOA and cloud computing share similar characteristics such as using services as basic blocks to build reliable and secure applications (Yi and Blake, 2010). Razavian and Lago report a systematic literature review of SOA migration approaches (Razavian and Lago, 2015). Their key goal of the survey is to identify commonalities and differences between 75 identified approaches and propose a reference model of typical activities that are carried out for the legacy to SOA migration. From the knowledge management point of view, this reference model includes typical knowledge that shapes a process of evolving legacy applications to service-based applications. Khadka et. al. provide a historic review ofmethodologies for the legacy to SOA migration (Khadka et al., 2013). The objectives of this review are (i) to reach a broad understanding of existing process models for legacy evolution to SOA (ii) identify available techniques to perform migration activities, and (iii) identify the existing issues and possible directions for future research. Through evaluating 121 primary studies using an evaluation framework, inspired from three traditional reengineering methodologies namely Butterfly (Bisbal et al., 1997), Renaissance (Warren and Ransom, 2002), and Architecture-Driven Modernization (ADM) (Khusidman and Ulrich, 2007), the authors conclude that there is still a lack of adequate automation level and techniques for determining the decomposability of legacy applications, investigating organisational perspective of migration, and postmortem reports on after-migration experience. In another attempt, Lane et. al. present a survey of process models to develop service-based applications with a skew towards dynamic adaptation (Lane and Richardson, 2011). Subsequently, they developed a meta-model giving an overarching view of development processes of 75 identified methodologies. On the basis of evaluation results using this meta-model, they found that increasing the automation of development process using model-driven development techniques is the most common theme in the reviewed methodologies. They also argued that existing methodologies suffer from a lack of real empirical validation. Even though the above-mentioned surveys are helpful, they are silent to address the cloudcentric challenges stated in Section 2.2.

The survey provided in the current study is different from the existing reviews in three salient aspects. Firstly,

this survey limits its focus on all extant approaches proposing a (complete or partial) migration process model or framework for the cloud migration, and hence is more specific than the above-mentioned surveys. None of the reviewed surveys (see Table II) provides an in-depth discussion on the features and migration activities proposed in the existing approaches as well as useful experience of applying these approaches in practice. Secondly, this survey provides an in-depth analysis of existing approaches through an evaluation framework, which encompasses 28 criteria classified into two dimensions i.e. generic and cloud-specific ones. The proposed framework has been derived through an extensive literature review and validated through a Web-based questionnaire survey of 104 experts from academia and experts in the field of cloud computing. Since all related surveys fail to consider the important evaluation criteria that the proposed framework includes, the evaluation framework can be considered as an important contribution of the current study. The characterisation framework of (Jamshidi et al., 2013) does not include any generic criteria as offered by our proposed evaluation framework. For the cloud-specific dimension, although 11 criteria have been referred by their framework, there is no elaboration on assessment of approaches. Thirdly, given our different focus, none of the related work covers the papers that this paper reviews. We found that only 5 out of our 43 reviewed papers were covered by (Jamshidi et al., 2013). Finally, this survey considers different and recently published approaches that are not covered in the other surveys. With respect to this, this survey can be viewed as complementary one to the above-mentioned surveys through investigating different and recent approaches.

As with all new areas of study, an etymological analysis is instructive. This is first undertaken in this section to give some clarity as to what a cloud migration methodology might mean in the context of cloud computing. This section then identifies technical and organisational concerns of such a methodology and provides a review of surveys related efforts.

2.1 ETYMOLOGY —Cloud migration methodology. In software engineering (SE) a software development methodology can be defined as a systematic way of doing things in a particular discipline (Gonzalez-Perez and Henderson-Sellers, 2008). Another definition can be borrowed from Avison and Fitzgerald: a recommended collection of phases, procedures, rules, techniques, tools, documentation, management and training used to develop a system (Avison and Fitzgerald, 2003). A methodology organises the coordination of development team members and integration project activities. It defines when a certain activity, which contains sequence and input/output artefacts, should be carried out. Migration of legacy applications to the cloud signifies that the organisation has already in place existing software applications earmarked to take advantages of cloud services. A common understanding of the term cloud migration methodology, as offered by (Chauhan and Babar, 2012), is the re-engineering process of legacy applications for becoming cloud-enabled. That is, migration to cloud is a kind of software reengineering where the target application will be able to interact or become integrated with cloud services. Another definition, offered by Andrikopoulos, views the cloud migration process as a set of architectural adaptations required to ensure a legacy application becoming cloud-compliant (Andrikopoulos et al., 2013). Similarly, Kwon et al. pose the term cloud refactoring in which code transformation mechanisms are used to integrate legacy applications and cloud services (Kwon and Tilevich, 2014). Another yet broader and workable definition, which covers both technical and nontechnical aspects of the cloud migration is suggested by (Pahl et al., 2013) as: A cloud migration process is a set of migration activities carried to support an end-to-end cloud migration. Cloud migration processes define a comprehensive perspective, capturing business and technical concerns. Stakeholders with different backgrounds are involved. One can envisage a cloud migration methodology as an extended traditional software development methodology to enhance its capability to support cloud computing. —Legacy Application. This paper focuses on approaches addressing the migration of legacy applications to cloud environments. As such, we also analyse the term legacy. In software engineering literature many definitions can be found for the term legacy applications. One of the earliest definitions is the following: large software systems that we don't know how to cope with but there are vital to our organization (Bennett, 1995). Similarly, Stone braker mentions that a legacy application is any system that significantly

resists modification and evolution (Brodie and Stonebraker, 1995). Sneed states they are information systems that have been in use for years (Sneed, 2006). Others emphasise technological aspects. E.g1, Stone distinguishes legacies as those that are not Internet-dependent (Stone, 2001). E.g2, Dedeke defines it as an aggregate package of software and hardware solutions whose languages, standards, codes, and technologies belong to a prior generation or era of innovation (Dedeke, 2012).

3. PROBLEM STATEMENT:

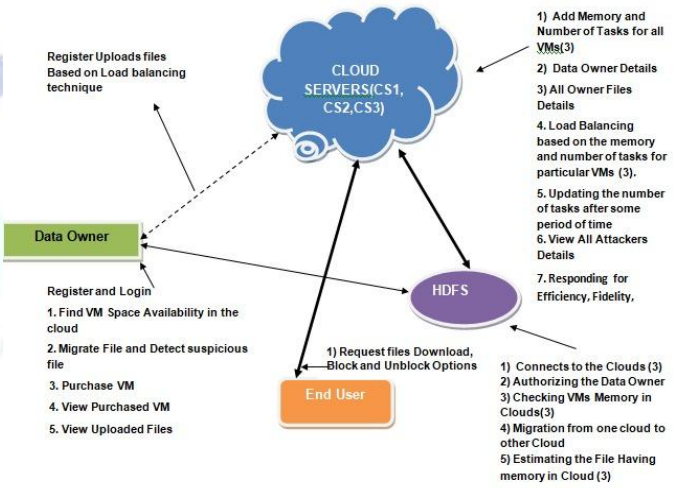
Data location in the cloud environment has been recognizedas an important factor in providing users with assurance of data security and privacy [13]. There have been some effortson the research problem of data placement control in cloudstorage systems. Peterson et al. [14] defined the notion of"data sovereignty" and proposed a MAC-based proof of datapossession (PDP) technique to authenticate the geographical locations of data stored in the cloud. Benson et al. [15]addressed the problem of determining the physical locations ofdata stored in geographically distributed data centers, by usingpassive distance measurement and linear regression predictivemodel to estimate in which data center the data is stored. Later,Gondree and Peterson [16] proposed a general framework,named constraint-based data geo-location (CBDG), that bindslatency-based geo-location techniques with a probabilisticPDP, based on the previous solutions in [14], [15]. In addition,Watson et al. [17] considered the case of collusion betweenmalicious service providers and suggested a proof of location(PoL) scheme that deployed trusted landmarks to verify the theexistence of a file on a host using proof of retrievability (PoR)protocol. In [18], [19], PoR was also adopted with a timebaseddistance-bounding protocol to provide strong geographical location assurance.

Instead of verifying file locations afterwards, another commonapproach is to require users to encrypt their data beforeuploading it to the cloud [20]. The rationale behind is the cloud does not have the original plain-text data, userswould have fewer concerns on data location. This approach,however, imposes a large computational burden on the usersand it renders the data hard to index and analyze on cloudpremises.

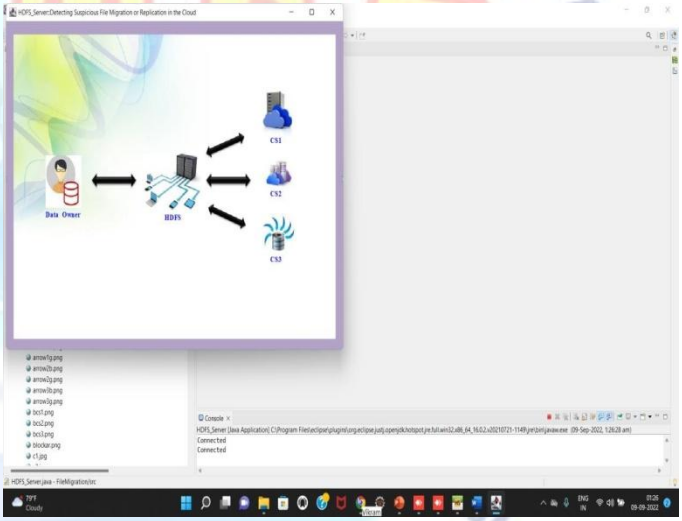
DISADVANTAGES

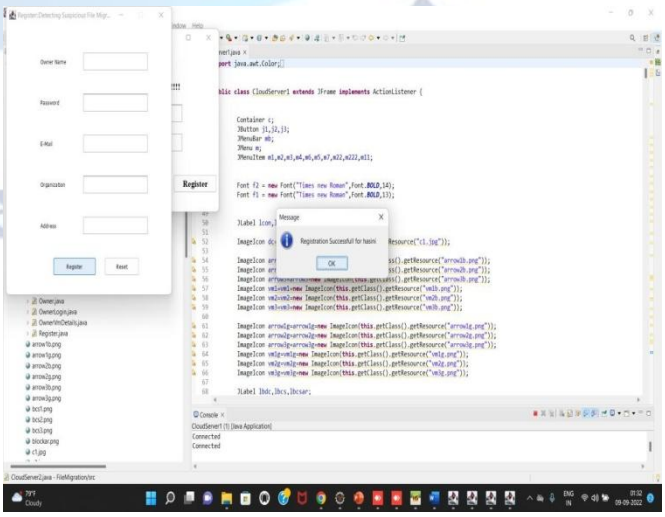
- In the existing work, the system doesn't have Location-aware File Loading techniques.
- There is no authentication and key agreement for Location-aware Load Balancing.

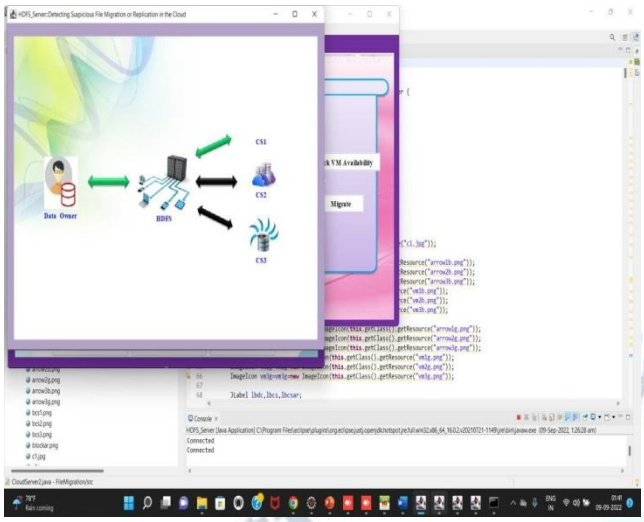
4. ARCHITECTURE:

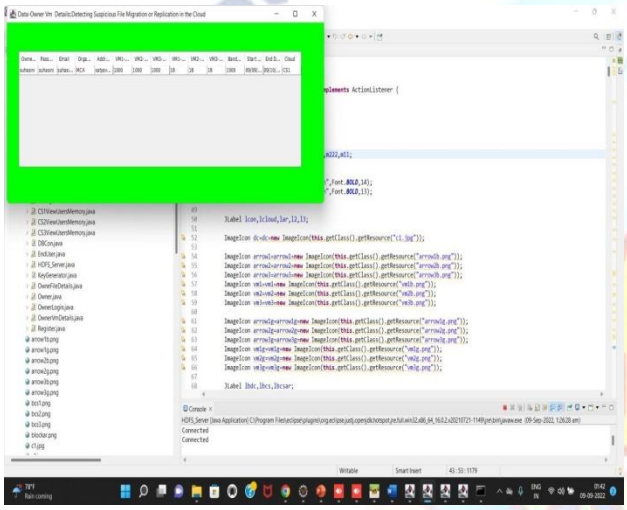


5. RESULT:









6. CONCLUSION:

In this study, we develop a novelLAST-HDFS system on top of the current HDFS to handle the problem of data placement control in the cloud. More importantly, LAST-HDFS assures that the location policy is enforced regardless of data replication and load balancing mechanisms that may impair policy compliance. Policy-driven fileloading enables location-aware storage on cloud sites. In particular, an effective LP-tree and Legal File Transfer graph were created to help allocate files with similar location preferences to the best cloud nodes, increasing the likelihood of spotting illegal file transfers. Both a large-scale simulated cloud environment and a real cloud testbed were used for ourthorough experimental research. The outcomes of these experiments demonstrated the usefulness and efficiency of the proposed LAST-HDFS system.

In the future, we plan to take into account more complicatedpolicies to capture other privacy requirements other than thelocation. We will adopt more sophisticated policy analysisalgorithm [21] and compute the integrated policy as the representativepolicy [22] at each node to help speed up the policycomparison and selection of nodes for the newly uploadedfiles. Moreover, we also plan to leverage Intel SGX technologyto secure socket monitors from being compromised.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Amazon, "Aws global infrastructure," in https://aws.amazon.com/aboutaws/ global-infrastructure/, 2017.

[2] C. Metz, "Facebook tackles (really) big data with project prism," in https://www.wired.com/2012/08/facebook-prism/, 2012.

[3] K. V. SHVACHKO, Y. Aahlad, J. Sundar, and P. Jeliazkov, "Geographically-distributed file system using coordinated namespace replication," in https://www.google.com/patents/WO2015153045A1?cl=zh, 2014.

[4] C. Liao, A. Squicciarini, and L. Dan, "Last-hdfs: Location-aware storage technique for hadoop distributed file system," in IEEE International Conference on Cloud Computing (CLOUD), 2016.

[5] N. Paladi and A. Michalas, ""one of our hosts in another country": Challenges of data geolocation in cloud storage," in International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 2014, pp. 1–6.

[6] Z. N. Peterson, M. Gondree, and R. Beverly, "A position paper on data sovereignty: The importance of geolocating data in the cloud." in HotCloud, 2011.

[7] A. Squicciarini, D. Lin, S. Sundareswaran, and J. Li, "Policy driven node selection in mapreduce," in 10th International Conference on Security and Privacy in Communication Networks (SecureComm), 2014.

[8] J. Li, A. Squicciarini, D. Lin, S. Liang, and C. Jia, "Secloc: Securing location-sensitive storage in the cloud," in ACM symposium on access control models and technologies (SACMAT), 2015.

[9] E. Order, "Presidential executive order on strengthening the cybersecurity of federal networks and critical infrastructure," in https://www.whitehouse.gov/the-press-office/2017/05/11/presidentialexecutive-order-strengthening-cybersecurity-federal, 2017.

[10] "Hdfs architecture," http://hadoop.apache.org/docs/stable/ hadoop-project-dist/hadoop-hdfs/HdfsDesign.html.

[11] R. Miller, "Inside amazon cloud computing infrastructure," in http://datacenterfrontier.com/inside-amazon-cloud-computinginfrastructure/, 2015.

[12] T. Bujlow, K. Balachandran, S. L. Hald, M. T. Riaz, and J. M. Pedersen,
"Volunteer-based system for research on the internet traffic," Telfor Journal, vol. 4, no. 1, pp. 2–7, 2012.
[13] M. Geist, "Location matters up in the cloud," http://www.thestar.com/
business/2010/12/04/geist location matters up in the cloud.html.
[14] Z. N. Peterson, M. Gondree, and R. Beverly, "A position paper on data sovereignty: the importance of geolocating data in the cloud," in Proceedings of the 8th USENIX conference on Networked systems design
and implementation, 2011.
[15] K. Benson, R. Dowsley, and H. Shacham, "Do you know where your
cloud files are?" in Proceedings of the 3rd ACM workshop on Cloud computing security workshop. ACM, 2011, pp. 73–82.
[16] M. Gondree and Z. N. Peterson, "Geolocation of data in the cloud," in Proceedings of the third ACM conference on Data and application security and privacy. ACM, 2013, pp. 25–36.
[17] G. J. Watson, R. Safavi-Naini, M. Alimomeni, M. E. Locasto, and S. Narayan, "Lost: location based storage," in Proceedings of the 2012 ACM Workshop on Cloud computing security workshop. ACM, 2012, pp. 59–70.
[18] A. Albeshri, C. Boyd, and J. G. Nieto, "Geoproof: proofs of geographic
location for cloud computing environment," in Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on.
IEEE, 2012, pp. 506–514.
[19] A. Albeshri, C. Boyd, and J. G. Nieto, "Enhanced geoproof: improved
geographic assurance for data in the cloud," International Journal of Information Security, vol. 13, no. 2, pp. 191–198, 2014. [20] A. Michalas and K. Y. Yigzaw, "Locless: Do you really care where your
cloud files are?" ACM/IEEE, 2016.
[21] D. Lin, P. Rao, R. Ferrini, E. Bertino, and J. Lobo, "A similarity measure for comparing XACML policies," IEEE Trans. Knowl. Data Eng., vol. 25, no. 9, pp. 1946–1959, 2013.
[22] P. Rao, D. Lin, E. Bertino, N. Li, and J. Lobo, "Fine-grained integration
of access control policies," Computers & Security, vol. 30, no. 2-3, pp. 91–107, 2011.

Authors Biography

K.Hruthpadma Suhasini currently pursuing MCA in SVKP & Dr.K.S Raju Arts & Science College affiliated to Adikavi Nannayya University, Rajamahendravaram, Her research interests include C & JAVA , Data Structures, Web Technologies, Operating Systems, Data Science and Artificial Intelligence.

B.N.Srinivasa Gupta is working as Associate Professor in SVKP &Dr K S Raju Arts & Science College, Penugonda, A.P. He received Masters Degree in Computer Applications from Andhra University and Computer Science & Engineering from Jawaharlal Nehru Technological University Kakinada (JNTUK),Kakinada, India. His research interests include Data Mining, Cyber Security, Artificial Intelligence