IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 2, March-April 2025

Impact Factor: 8.152

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

Cloud File Migration and Threat Detection

Sudhavali Adusumalli [1], Kagga Yojitha Bindu Madhavi [2], Angadala Durga Bhavani [3], Bommana Nikhil Venkata Sai [4], Banavathu Shankar [5]

Assistant Professor, Department of CSE, Eluru College of Engineering & Technology, Eluru, India [1]

B. Tech Student, Department of CSE, Eluru College of Engineering & Technology, Eluru, India[2,3,4,5]

ABSTRACT: There has been a prolific rise in the popularity of cloud storage in recent years. While cloud storage offers many advantages, such as flexibility and convenience, users are typically unable to tell or control the actual locations of their data. This limitation may affect users' confidence and trust in the storage provider, or even render the cloud unsuitable for storing data with strict location requirements. To address this issue, we propose a system called LAST-HDFS, which integrates Location-Aware Storage Technique (LAST) into the open-source Hadoop Distributed File System (HDFS). The LAST-HDFS system enforces location-aware file allocations and continuously monitors file transfers to detect potentially illegal transfers in the cloud.

Illegal transfers here refer to attempts to move sensitive data outside the ("legal") boundaries specified 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 being 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.

KEYWORDS: Machine Learning, Cloud Storage, Detection and Location-Aware Storage Technique.

I. INTRODUCTION

Computing firms are no longer the only consumers of cloud storage and cloud computing, but rather average businesses, and even end-users are taking advantage of the immense capabilities that cloud services can provide. While enjoying the flexibility and convenience brought by cloud storage, cloud users release control over their data, and particularly are often unable to locate their actual data; this could be in-state, in-country, or even out-of-country. Lack of location control may cause privacy breaches for cloud users (e.g., hospitals) who store sensitive data (e.g., medical records) that are governed by laws to remain within certain geographic boundaries and borders. Another situation where this problem arises is with governmental entities that require all data to be stored in the same country where the government operates; this challenge has seen difficulties with cloud service providers (CSPs) quietly moving data out of the country or being bought out by foreign companies.

For example, Canadian laws demand that personal identifiable data must be stored in Canada. However, large cloud infrastructure like the Amazon Cloud has more than 40 zones distributed all over the world [1], which makes it very challenging to provide guaranteed adherence to regulatory compliance. Even Hadoop, which historically has been managed as a geographically confined distributed file system, is now deployed on a large scale across different regions (see Facebook Prism or recent patent).

To date, various tools have been proposed to help users verify the exact location of data stored in the cloud, with emphasis on post-allocation compliance. However, recent work has acknowledged the importance of a proactive location control for data placement consistent with adopters' location requirements, to allow users to have stronger control over their data and to guarantee the location where the data is stored. In this work, we infiltrate one of the most widely adopted cloud data storage systems–Hadoop Distributed File System (HDFS), and design an enhanced HDFS system, called LAST-HDFS. The LAST-HDFS extends HDFS' capabilities to achieve location-aware file allocations and file transfer monitoring. Specifically, LAST-HDFS provides the following new functions:

i. consistently enforces a location-aware data loading and storage by assigning datanodes according to user-

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

specified privacy policies;

ii. actively tracks and dynamically corrects possible data migration (due to balancing or data replication needs) within the cluster that might violate data placement policies;

1.1 MOTIVATION

Many people and businesses store files in the cloud. Moving these files can be hard and risky. We need ways to move files easily and safely. Cloud storage also faces security threats. This project makes moving files simpler. It also finds and stops threats to those files. We want to make sure files are safe and easy to use in the cloud. This helps people use cloud storage without worrying about losing data or getting attacked**.**

1.2 PROBLEM DEFINITION

Organizations face significant challenges when moving files to the cloud: transfers are often slow and risky, and existing tools don't adequately protect data from new cyber threats. Current systems handle migration and security separately, creating gaps and requiring manual work, which leads to potential data loss, security breaches, and difficulty in managing cloud files effectively. This project aims to solve these problems by creating a single system that makes cloud file migration easier and more secure.

1.3 OBJECTIVE OF THE PROJECT

The primary objective of this project is to develop a comprehensive system that streamlines cloud file migration while simultaneously enhancing security through advanced threat detection. This involves creating an intelligent platform that automates migration processes, optimizes resource utilization, and ensures data integrity during transfers. Furthermore, the system will implement real-time threat analysis and proactive security measures to detect and mitigate malware, unauthorized access, and other security risks, ultimately providing organizations with a secure, efficient, and compliant cloud file management solution.

II. LITERATURE SURVEY

2.1 AWS global infrastructure

The AWS Cloud infrastructure is built around AWS Regions and Availability Zones. An AWS Region is a physical location in the world where we have multiple Availability Zones. Availability Zones consist of one or more discrete data centers, each with redundant power, networking, and connectivity, housed in separate facilities. These Availability Zones offer you the ability to operate production applications and databases that are more highly available, fault-tolerant, and scalable than would be possible from a single data center. For the latest information on the AWS Cloud Availability Zones and AWS Regions.

2.2 Geographically-distributed file system using coordinated namespace replication

A cluster of nodes implements a single distributed file system, comprises at least first and second data centers, and a coordination engine process. The first data center may comprise first DataNodes configured to store data blocks of client files, and first NameNodes configured to update the state of a namespace of the cluster. The second data center, geographically remote from and coupled to the first data center by a wide area network, may comprise second DataNodes configured to store data blocks of client files, and second NameNodes configured to update the state of the namespace. The first and second NameNodes are configured to update the state of the namespace in response to data blocks being written to the DataNodes. The coordination engine process spans the first and second NameNodes and coordinates updates to the namespace stored such that the state thereof is maintained consistent across the first and second data centers.

2.3 Last-hdfs: Location-aware storage technique for Hadoop distributed file system

Enabled by the state-of-the-art cloud computing technologies, cloud storage has gained increasing popularity in recent years. Despite the benefit of flexible and reliable data access offered by such services, users have to bear with the fact of not knowing the whereabouts of their data. The lack of knowledge and control of the physical locations of data could raise legal and regulatory issues, especially for certain sensitive data that is governed by laws to remain within certain geographic boundaries and borders. In this paper, we study the problem of data placement control within distributed file systems supporting cloud storage. Particularly, we consider the open-source Hadoop file system (HDFS) as the underlying architecture, and propose a location-aware cloud storage system, named LAST-HDFS, to support and enforce location-aware storage in HDFS-based clusters. In addition, it also includes a monitoring system deployed at individual hosts to oversee and detect potential data placement violations due to the existence of malicious datanodes.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

We carried out an extensive experimental evaluation in a real cloud environment that demonstrates the effectiveness and efficiency of our proposed system.

2.4 One of our hosts in another country

The physical location of data in cloud storage is an increasingly urgent problem. In a short time, it has evolved from the concern of a few regulated businesses to an important consideration for many cloud storage users. One of the characteristics of cloud storage is fluid transfer of data both within and among the data centres of a cloud provider. However, this has weakened the guarantees for control over data replicas, protection of data in transit, and physical location of data. This paper addresses the lack of reliable solutions for data placement control in cloud storage systems. We analyse the currently available solutions and identify their shortcomings. Furthermore, we describe a high-level architecture for a trusted, geolocation-based mechanism for data placement control in distributed cloud storage systems, which are the basis of an ongoing work to define the detailed protocol and a prototype of such a solution. This mechanism aims to provide granular control over the capabilities of tenants to access data placed on geographically dispersed storage units comprising the cloud storage.

2.5 A position paper on data sovereignty: The importance of geolocating data in the cloud.

In this paper we define the problem and scope of data sovereignty - the coupling of stored data authenticity and geographical location in the cloud. Establishing sovereignty is an especially important concern amid legal and policy constraints when data and resources are virtualized and widely distributed. We identify the key challenges that need to be solved to achieve an effective and uncheatable solution, as well as propose an initial technique for data sovereignty.

2.6 Policy-driven node selection in MapReduce

The MapReduce framework has been widely adopted for processing Big Data in the cloud. While efficient, MapReduce offers very complicated (if any) means for users to request nodes that satisfy certain security and privacy requirements to process their data. In this paper, we propose a novel approach to seamlessly integrate node selection control to the MapReduce framework for increasing data security. We define a succinct yet expressive policy language for MapReduce environments, according to which users can specify their security and privacy concerns over their data. Then, we propose corresponding data preprocessing techniques and node verification protocols to achieve strong policy enforcement. Our experimental study demonstrates that, compared to the traditional MapReduce framework, our policy control mechanism allows for achieving data privacy without introducing significant overhead.

2.7 Secloc: Securing location-sensitive storage in the cloud

Cloud computing offers a wide array of storage services. While enjoying the benefits of flexibility, scalability and reliability brought by the cloud storage, cloud users also face the risk of losing control of their own data, in partly because they do not know where their data is actually stored. This raises a number of security and privacy concerns regarding one's sensitive data such as health records. For example, according to Canadian laws, data related to personal identifiable information must be stored within Canada. Nevertheless, in contrast to the urgent demands, privacy requirements regarding to cloud storage locations have not been well investigated in the current cloud computing market, fostering security and privacy concerns among potential adopters. Aiming at addressing this emerging critical issue, we propose a novel secure location-sensitive storage framework, called SecLoc, which offers protection for cloud users' data following the storage location restrictions, with minimum management overhead to existing cloud storage services. We conduct security analysis, complexity analysis and experimental evaluation on the proposed SecLoc system. Our results demonstrate both the effectiveness and efficiency of our mechanism.

2.8 HDFS architecture

HDFS is an open-source component of the Apache Software Foundation that manages data. HDFS has scalability, availability, and replication as key features. Name nodes, secondary name nodes, data nodes, checkpoint nodes, backup nodes, and blocks all make up the architecture of HDFS. HDFS is fault-tolerant and is replicated. Files are distributed across the cluster systems using the Name Node and Data Nodes. The primary difference between Hadoop and Apache HBase is that Apache HBase is a non-relational database, and Apache Hadoop is a non-relational data store.

2.9 Location matters in the cloud

For certain types of sensitive data (such as health records), it is important to know the geographic location of the file, e.g., that it is stored on servers within the USA. This is particularly important for determining applicable laws and regulations. In this paper, we discuss the problem of verifying the location of files within distributed file storage systems such as the cloud. We consider a general setup for a distributed storage system and show that verifying location when such a system is fully malicious is impossible. We then make plausible assumptions about the behavior

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

of the system and provide a formal definition for Proofs of Location (PoL) in our setting. We show that secure and efficient PoL schemes can be constructed by using a geolocation scheme and a Proof of Retrievability (PoR) scheme with a new added property that we call re-coding, which is of independent interest.

III. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Currently, cloud file migration and threat detection often exist as distinct, fragmented systems. Migration tools primarily handle basic file transfers, lacking intelligent automation, real-time validation, and optimized bandwidth usage. Security solutions rely heavily on signature-based detection and perimeter security, struggling to address evolving cloud-native threats and insider risks. Integration between migration and security is minimal, leading to manual intervention, delayed threat response, and limited visibility into file activity. This disjointed approach creates operational inefficiencies and security gaps, leaving organizations vulnerable to data loss and security breaches during and after cloud migration.

3.1.1 Disadvantages of the Existing System

- Lack of Integrated Security
- Inefficient Migration Processes
- Limited Threat Detection
- Manual Intervention Required
- Insufficient Visibility and Control

3.2 PROPOSED SYSTEM

The proposed system aims to create a unified platform that seamlessly integrates cloud file migration with advanced threat detection, addressing the shortcomings of existing fragmented solutions. It will feature intelligent migration capabilities, including automated policy-driven transfers, real-time data validation, and optimized bandwidth utilization, significantly reducing migration time and minimizing data loss. Simultaneously, the system will incorporate behavioral analytics and machine learning to detect anomalous file access patterns and zero-day threats, providing proactive security measures beyond traditional signature-based detection.

Furthermore, the proposed system will prioritize comprehensive security and compliance, offering end-to-end encryption, granular access control, and automated reporting. It will provide a centralized dashboard for real-time monitoring of migration progress, threat detection, and security events, offering organizations enhanced visibility and control over their cloud file activities. Integration with SIEM/SOAR systems will enable automated threat response and streamlined incident management. By consolidating migration and security functionalities into a cohesive platform, the proposed system will empower organizations to securely and efficiently manage their cloud file lifecycle.

3.2.1 Advantages of the Proposed System

- Enhanced Security
- Streamlined Migration
- Proactive Threat Response
- Improved Visibility and Control

IV. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

- The DFD is also called a bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data generated by this system.
- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
- DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

- DFD is also known as a bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

4.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general- purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to or associated with, UML. The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

UML was created as a result of the chaos revolving around software development and documentation. In the 1990s, there were several different ways to represent and document software systems.
The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

4.2a GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of object oriented tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

V. RESULTS

The following figures present the sequence of screenshots of the results.

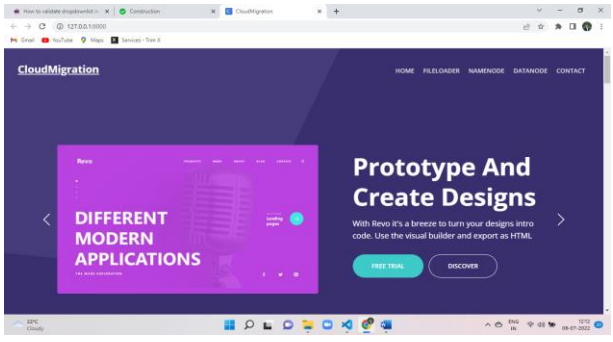

Fig 2a: Home Page

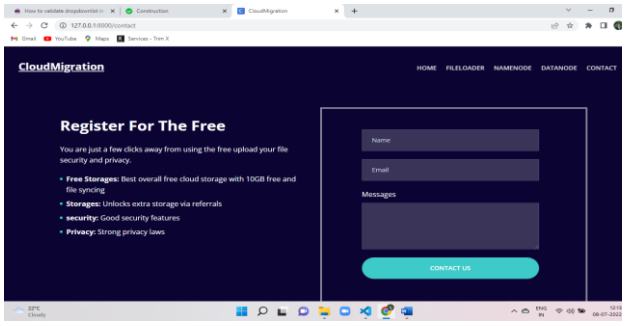

Fig 2b: CONTACT for Communication page

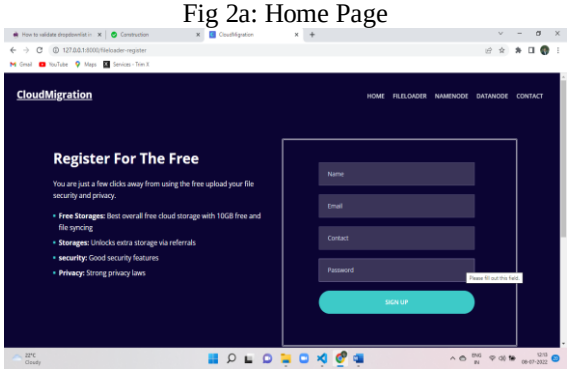

Fig 2c: "FILELOADER" to create Account

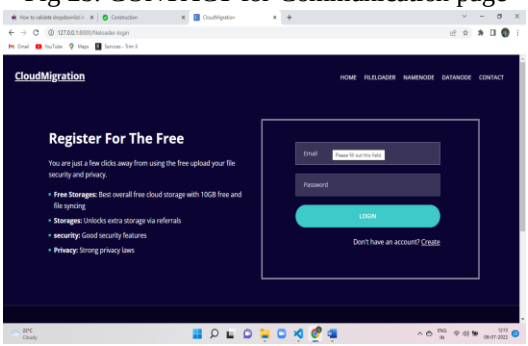

Fig 2d: login with your credentials

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

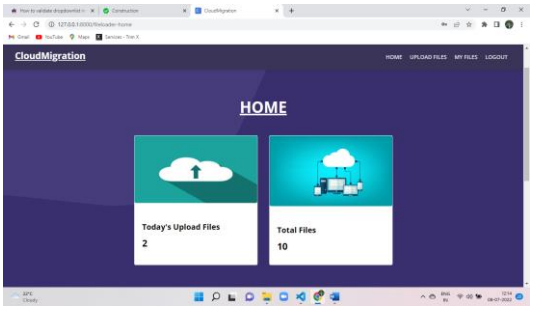

Fig 2e: "UPLOAD FILES" to upload your file.

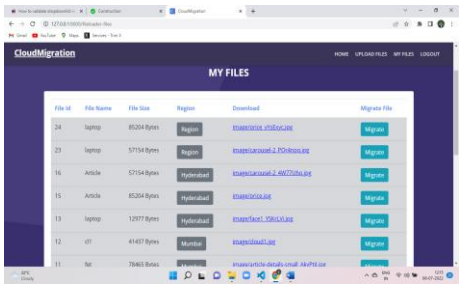

Fig 2f: "Migrate" to migrate your file

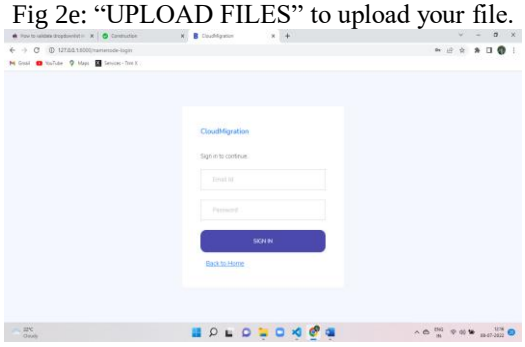

Fig 2g: "NAMENODE" for admin login

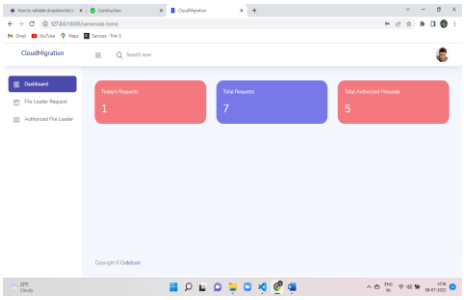

Fig 2h: Home Page of the Admin.

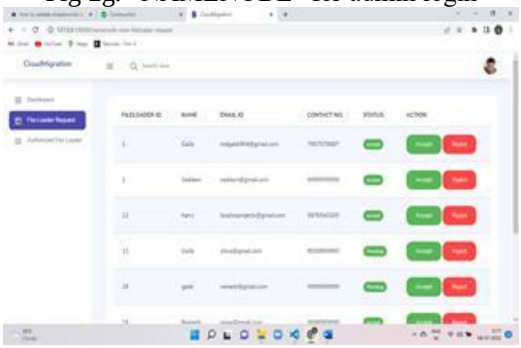

Fig 2i:c "File Loader Request"

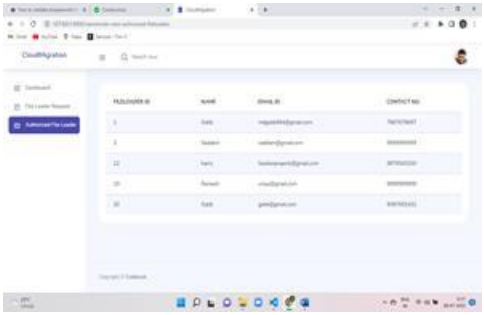

Fig 2j: "Authorized File Loader"

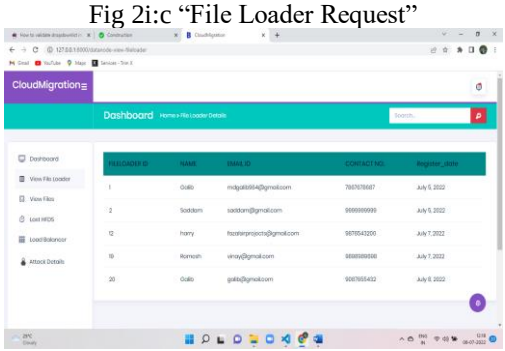

Fig 2k: View File Loader Details

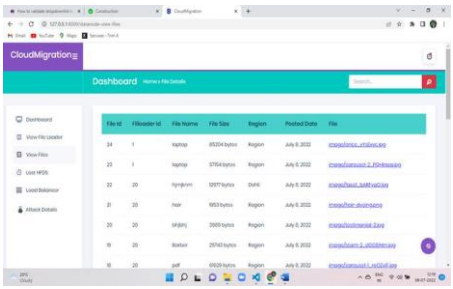

Fig 2l: View the Files Uploaded by the User

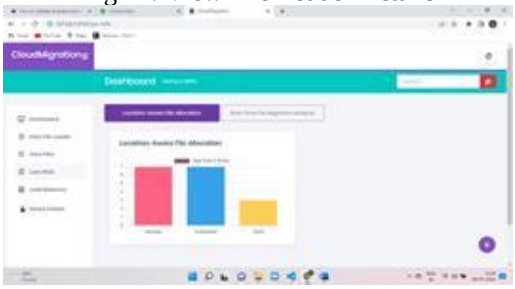

Fig 2m: Bar Graph showing the number of files in each location.

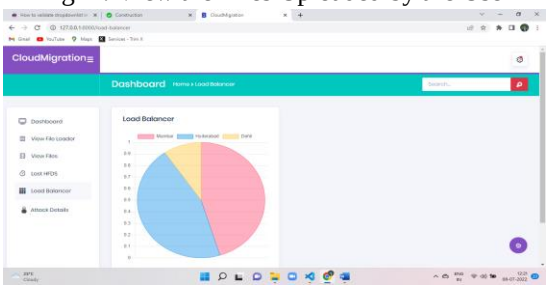

Fig 2n: check load of files in each region

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

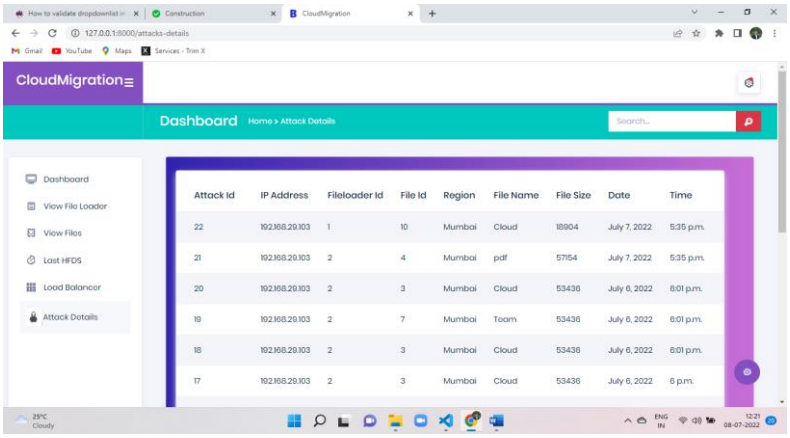

Fig 2o: View the Attack Details of the files.

VI. CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

In this paper, we build, on top of the existing HDFS, a novel LAST-HDFS system to address the data placement control problem in the cloud. LAST-HDFS supports policy-driven file loading that enables location-aware storage in cloud sites. More importantly, it also ensures that the location policy is enforced regardless of data replication and load balancing processes that may affect policy compliance. Specifically, an efficient LP-tree and Legal File Transfer graph were designed to help optimally allocate files with similar location preferences to the most suitable cloud nodes which in turn enhance the chance of detecting illegal file transfers. We have conducted extensive experimental studies in both a real cloud testbed and a large-scale simulated cloud environment. Our experimental results have shown the effectiveness and efficiency of the proposed LAST-HDFS system.

In the future, we plan to take into account more complicated policies to capture other privacy requirements other than the location. We will adopt more sophisticated policy analysis algorithm and compute the integrated policy as the representative policy at each node to help speed up the policy comparison and selection of nodes for the newly uploaded files. Moreover, we also plan to leverage Intel SGX technology to secure socket monitors from being compromised.

6.2 FUTURE WORK

Future work for "Cloud File Migration and Thread Detection" should prioritize enhancing migration capabilities through intelligent policies, cross-cloud interoperability, delta migration, and performance optimizations like parallel processing and serverless functions. Simultaneously, advanced threat detection can be achieved by implementing behavioral analysis, real-time threat response, and deep learning for malware detection. Integrating with SIEM/SOAR systems and providing robust forensics and auditing tools will strengthen security. Further improvements include end-to-end encryption, automated compliance, and user-friendly interfaces with automated reporting, API integration, and simulation tools.

Additionally, the project should focus on expanding security and compliance features, including granular access control, multi-factor authentication, and data sovereignty measures. User experience enhancements, such as intuitive interfaces, automated reporting, and API integration, are crucial. Finally, exploring edge computing integration for localized processing and optimizing for hybrid cloud environments will address modern deployment challenges, ensuring the project remains relevant and effective in the evolving cloud landscape.

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.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, March-April 2025 ||

DOI:10.15680/IJARETY.2025.1202052

[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] K. Oku, R. K. Vaddy, A. Yada, and R. K. Batchu, "Data Engineering Excellence: A Catalyst for Advanced Data Analytics in Modern Organizations," International Journal of Creative Research in Computer Technology and Design, vol. 6, no. 6, pp. 1–10, 2024.
[13] 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.
[14] M. Geist, "Location matters up in the cloud," http://www.thestar.com/ business/2010/12/04/geist location matters up in the cloud.html.
[15] 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.
[16] 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.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com