A NEW IDENTIFYING POTENTIALLY UNUSUSAL FILE TRANSFERS OR DUPLICATION'S IN CLOUD ENVIRONMENTS

Guide: Dr.BASKAR NANJAPPAN
Assistant Professor, Department of CSE**, Malla Reddy Engineering College for Women**, TELANGANA, India.
N.CHANDRIKA[1], M.SRIDEVI[2], P.THANMAY SREE[3], P.DIVYA[4]
B. Tech Pursuing , Department of CSE **, Malla Reddy Engineering College for Women**, TELANGANA, India.

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 cloud unsuitable for storing data with strict location requirements. To address this issue,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.

Index Terms*:* Hadoop Distributed File system,Location-Aware Storage Technique

INTRODUCTION
With the ever-increasing popularity of cloud computing, the demand for cloud storage has also increased exponentially. 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 the actual their 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 were this problem arises is with governmental entities that require all data to be stored in the same country
that the government operates in; this challenge has seen difficulties with cloud service providers (CSPs) quietly moving data out-of-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 in 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.

LITERATURE SURVEY

TITLE: "Aws global infrastructure"

ABSTRACT: 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,

TITLE: "Geographically-distributed file system using coordinated namespace replication,"

ABSTRACT: 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 a 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 responsive 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.

TITLE: Last-hdfs: Location-aware storage technique for hadoop distributed file system,"

ABSTRACT: Enabled by the state-of-the-art cloud computing technologies, cloud storage has gained increasing popularity in recent years. Despite of the benefit of flexible and reliable data access offered by such services, users have to bear with the fact of not actually 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 are 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. We carried out an extensive experimental evaluation in a real cloud environment that demonstrates the effectiveness and efficiency of our proposed system.

TITLE: "one of our hosts in another country"

ABSTRACT: 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 with respect

to 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 on-going 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.

EXISTING SYSTEM

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 cloud unsuitable for storing data with strict location requirements. While enjoying the flexibility and convenience brought by cloud storage, cloud users release control over their data, and particularly are often unable to locate the actual their 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 were this problem arises is with governmental entities that require all data to be stored in the same country that the government operates in; this challenge has seen difficulties with cloud service providers (CSPs) quietly moving data out-of-country or being bought out by foreign companies.

PROPOSED SYSTEM

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. 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 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.

SYSTEM ARCHITECTURE

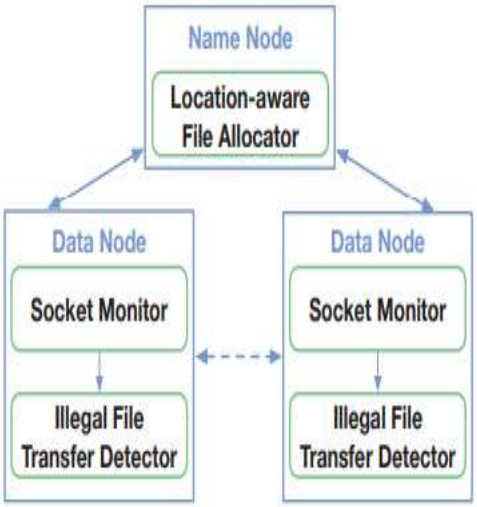
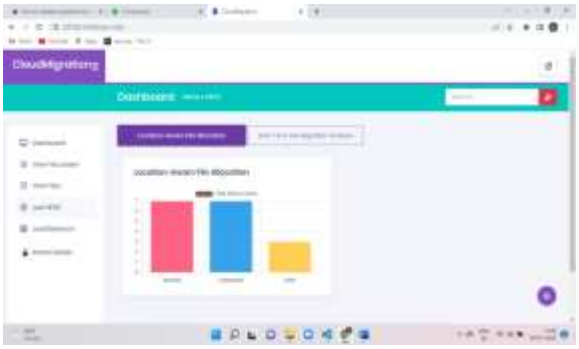


Fig. 1: Architecture of the Name Node and Data Nodes

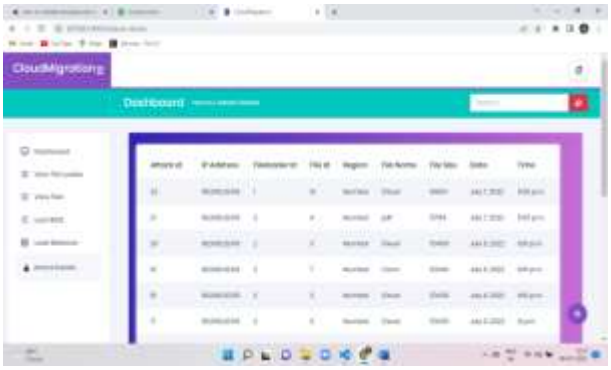
RESULTS



DATANODE VIEW REAL-TIME FILE MIGRATION ANALYSIS

DATANODE VIEW LOADBALANCER

DATANODE VIEW ATTACK DETAILS



ATTACKER VIEW FILE

CONCLUSION

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.

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.

AUTHOR

Dr.Baskar Nanjappan Professor Department of CSE MallaReddy Engineering College for Women,Hyderabad, baskarsrkv@gmail.com