International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST523102140

Load Balancing Using Oblivious RAM for Privacy-Preserving Access to Big Data in Cloud

P. B. Niranjane

Department of Computer Science & Engineering, Babasahe Naik College of Engineering, Pusad, Maharastra, India

ARTICLEINFO

Article History:

Accepted: 05 April 2023
Published: 28 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
885-889

ABSTRACT

In the era of big data, many users and companies start to move their data to cloud storage to simplify data management and reduce data maintenance cost. However, security and privacy issues become major concerns because third-party cloud service providers are not always trusty. Although data contents can be protected by encryption, the access patterns that contain important information are still exposed to clouds or malicious attackers. To overcome this problem, in this work Oblivious RAM (ORAM) is applied which aims to enable privacy preserving access to big data in cloud. After applying ORAM to big data in cloud the access load balancing problem arises. Iterative Load Balancing algorithm (ILB) is applied to achieve balanced load.

Keywords: Big data, Authentication Protocols, Privacy, Security, Encryption, Oblivious RAM, ILB

I. INTRODUCTION

Big data has emerged in various domains including science, engineering and commerce. For example, the amount of photos currently stored by Facebook is over 20 petabytes, and it continues to grow with 60 terabytes each week [1]. Sharing photos is one of Facebook's most popular features. To the date, users have uploaded over 65 billion photos making Facebook the biggest photo sharing website in the world. For each uploaded photo, Facebook generates and stores four images of different sizes, which translates to over 260 billion images and more than 20 petabytes of data. Users upload one billion new photos (˜60 terabytes) each week and Facebook serves over one million images per second at peak. As we expect these numbers to increase in the future, photo storage poses a significant challenge for Facebook's infrastructure.

In the era of big data, cloud becomes a perfect candidate for data storage by providing virtually unlimited storage that can be accessed over network. By outsourcing large volumes of data to cloud storage, such as Google Drive, Drop box and Amazon S3, users can simplify their data management and reduce data maintenance cost due to the pay-as-you-use model.

Because of security and privacy concerns, however, some users and companies still hesitate to move their data to cloud. Although encryption can protect the data confidentiality, it is insufficient because access patterns can also leak important information. For instance, over 80% of encrypted email queries can be identified according to access pattern [2]. Sensitive documents, however, need to be stored in encrypted format due to security concerns. But, encrypted storage makes it difficult to search on the stored documents. Therefore, this poses a major barrier towards selective retrieval of encrypted documents from the remote servers. Various protocols have been proposed for keyword search over encrypted data to address this issue. Most of the available protocols leak data access patterns due to efficiency reasons. Although, oblivious RAM based protocols can be used to hide data access patterns, such protocols are computationally intensive and do not scale well for real world datasets.

II. RELATED STUDIES

A. Big Data

The process of storing and analyzing data to make some sense for the organization is called Bigdata. In simple terms, data which is very large in size and yet growing exponentially with time is called as Bigdata. The term has been in use since the 1990s, with some giving credit to John Mashey for coining or at least making it popular. Bigdata usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Bigdata philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Bigdata requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale. Lot of confusion starts with the word Bigdata itself. For some group 1TB can be big, for others 10TB may be big. In simple words, we can say, "Bigdata is circumstances where the volume, velocity and variety of data go beyond an organization's storage or computation capacity for precise and well-timed decision making". Doug Laney characterize Big Data in terms of V 's.

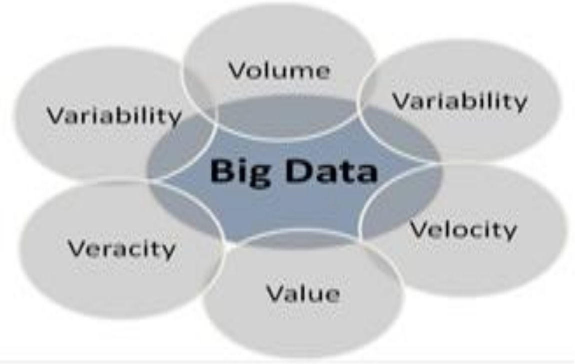


Figure 1: Characterization of Bigdata

B. Big Data and Cloud Computing

Big Data has emerged in the past few years as a new paradigm providing abundant data and opportunities to improve and/or enable research and decision-support applications with unprecedented value for digital earth applications including business, sciences and engineering. At the same time, Big Data presents challenges for digital earth to store, transport, process, mine and serve the data. Cloud computing provides fundamental support to address the challenges shared computing resources including computing, storage, networking and analytical software; the application of these resources has fostered impressive Big Data advancements Big Data poses unique challenges from several aspects including analysis, visualization, integration and architecture, due to the inherent high-dimensionality of geospatial data and the complex spatiotemporal relationships. To address Big Data challenges, a variety of methodologies, techniques and tools are identified to facilitate the transformation of data into value. Computing infrastructure, especially cloud computing, plays a significant role in information and knowledge extraction.

While the Big Data challenges can be tackled by many advanced technologies, such as HPC, cloud computing is the most elusive and important.

C. Privacy-Preserving In Cloud Storage

Cloud storage is often managed and maintained by cloud storage providers in the form of services, which is comprised of logical storage pools that interact directly with data, physical storage spanning across multiple servers, and physical environment. Users obtain storage capacity from the providers and operate their data through the public application programming interfaces (APIs). Many well-known providers have started their cloud storage services during the past few years, such as Microsoft SkyDrive, Amazon S3, and Google Drive. As the best infrastructure to accommodate big data, cloud storage has attracted a lot of attentions from both industry and academic. However, the security and privacy concerns arising from the natures of cloud storage are preventing users from subscribing to this service.

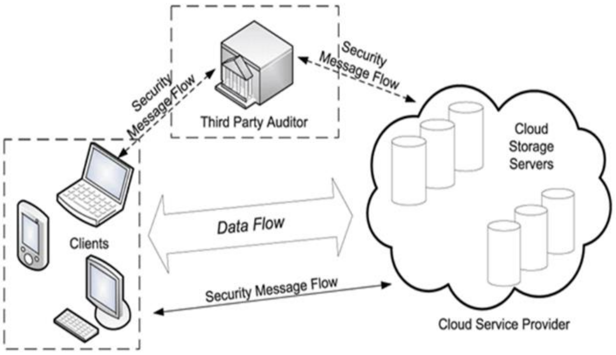


Figure 2: Privacy-Preserving In Cloud Storage

D. Security Techniques In Cloud Storage

With the techniques mentioned below the security and reliability of cloud storage can be improved.

1) RAID*:* RAID (Redundant Array of Inexpensive Disks) technique is integrated in HAIL [4]. RAID is a data storage virtualization technology that combines multiple physical disk drive components into a single logical unit for the purposes of data redundancy, performance improvement, or both.

2) HAIL: HAIL (High-Availability and Integrity Layer), is a distributed cryptographic system that permits a set of servers to prove to a client that a stored file is intact and retrievable. HAIL [5] strengthens, formally unifies, and streamlines distinct approaches from the cryptographic and distributed-systems communities. Proofs in HAIL are efficiently computable by servers and highly compact typically tens or hundreds of bytes, irrespective of file size. HAIL cryptographically verifies and reactively reallocates file shares. It is robust against an active, mobile adversary, i.e., one that may progressively corrupt the full set of servers.

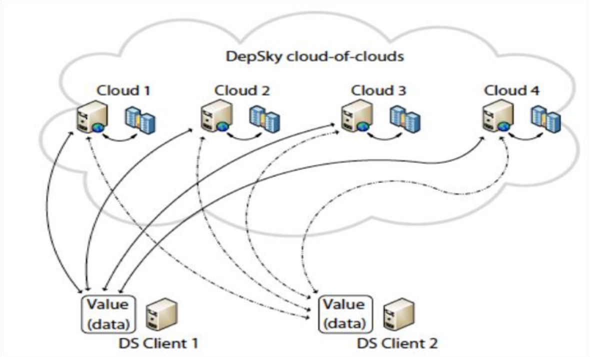


Figure 3: HAIL Architecture

3) IRIS: IRIS [4] has been proposed as an authenticated file system that lets enterprises store data in the cloud with resilience against potentially untrusted cloud providers. There are several proposals dealing with data availability by constructing distributed storage systems across several cloud sites.

4) DEPSKY: DEPSKY is a dependable and secure storage system that leverages the benefits of cloud computing by using a combination of diverse commercial clouds to build a cloud-of-clouds. In other words, DEPSKY is a virtual storage cloud, which is accessed by its users by invoking operations in several individual clouds.

III.PROPOSED WORK

This work proposed the system for load balancing for privacy preserving access to big data in cloud. Here clever cloud as cloud storage is used. ORAM algorithm used to hide access patterns and ILB used to balance the load. The proposed system consists of three sub modules are as follows:

1. Clever Cloud.
2. ORAM.
3. ILB

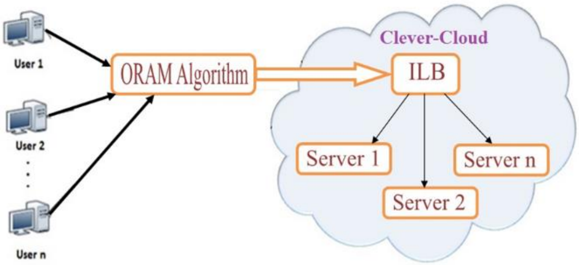


Figure 4: Load Balancing And Privacy Preserving Access To Big Data In Cloud

In this architecture the cloud is consist of huge amount of data or we can say “Big data”. The privacy-preserving data access to big data in an untrusted cloud by applying the ORAM algorithm is provided. In conjunction with encryption, ORAM-based solutions can hide not only data contents but also access patterns from third-party cloud service provider and malicious attackers. Then load balance problem will arrive on distributed file systems where we would apply ORAM. Finally, a low-complexity algorithm known as ILB (iterative load balancer) that can deal with large-scale problem instances with respect to big data is applied and the load distributes over the number of servers.

A. Clever Cloud

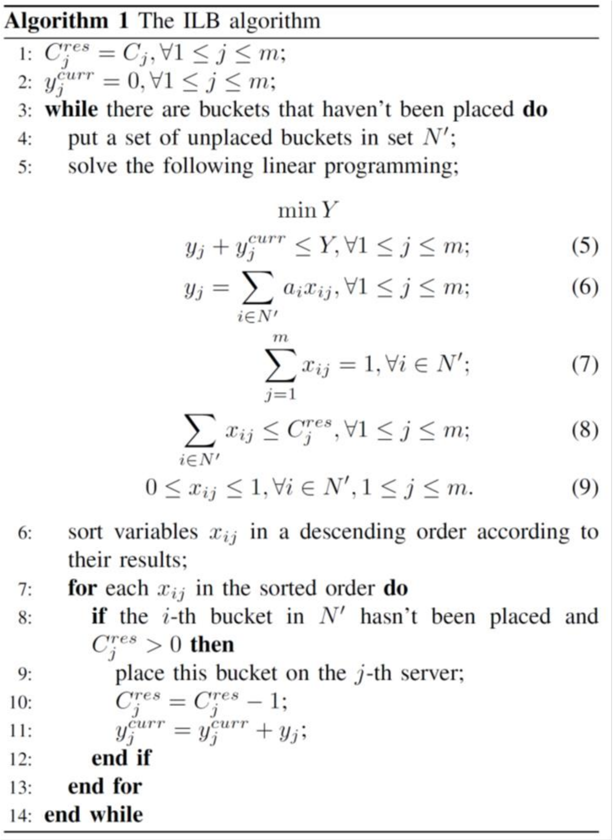
Clever coud is a Europe-based Paas company. Clever Cloud helps companies and IT professionals to achieve software delivery faster, reduce their feedback loop, focus on their core value and stop worrying about their hosting infrastructure by providing a solution for application sustainability. GitHub is where people build software. More than 27 million people use GitHub to discover, fork, and contribute to over 80 million projects.

B. ORAM

Recent ORAM work has explored optimizations of the classic Hierarchical Scheme [7], including the use of cuckoo hashing and Bloom filters. Williams et al. [9] have presented SR-ORAM as the first single-round-trip poly logarithmic time ORAM that requires only logarithmic client storage. Taking only a single round trip to perform a query. The presented ORAM algorithm is specifically applied to the scenario of cloud storage. A client is assumed that would like to store and retrieve its big data in cloud that is honest but curious. In other words, the cloud cannot tamper with or modify the data, but could learn information about the data. The data are divided into blocks, each of which is identified by a unique address.

C. ILB

ILB stands for Iterative Load Balancer. The typical use of load balancers are they used to increase capacity (concurrent users) and reliability of applications. They improve the overall performance of applications by decreasing the burden on servers associated with managing and maintaining application and network sessions, as well as by performing application-specific tasks. Scalable algorithms for the load balancing problem operate locally on the Nodes of the graph. They iteratively balance the load of a node with its neighbors until the whole network is globally balanced. The class of local iterative load balancing algorithms distinguishes between diffusion [6].



Algorithm 1 The ILB algorithm

```
$C_j^{res} = C_j, \forall 1 \le j \le m$;
$y_j^{curr} = 0, \forall 1 \le j \le m$;
while there are buckets that haven't been placed do
    put a set of unplaced buckets in set $N'$;
    solve the following linear programming;
        $\min Y$
        $y_j + y_j^{curr} \le Y, \forall 1 \le j \le m;$   (5)
        $y_j = \sum_{i \in N'} a_i x_{ij}, \forall 1 \le j \le m;$   (6)
        $\sum_{j=1}^{m} x_{ij} = 1, \forall i \in N';$   (7)
        $\sum_{i \in N'} x_{ij} \le C_j^{res}, \forall 1 \le j \le m;$   (8)
        $0 \le x_{ij} \le 1, \forall i \in N', 1 \le j \le m.$   (9)
    sort variables $x_{ij}$ in a descending order according to their results;
    for each $x_{ij}$ in the sorted order do
        if the $i$-th bucket in $N'$ hasn't been placed and $C_j^{res} > 0$ then
            place this bucket on the $j$-th server;
            $C_j^{res} = C_j^{res} - 1$;
            $y_j^{curr} = y_j^{curr} + y_j$;
        end if
    end for
end while
```

The quality of a balancing algorithm can be measured in terms of number of iterations it requires to reach a balanced state and in terms of the amount of load moved over the edge of the graph. The ILB algorithm is given above.

The system is implemented using Java as programming language and NetBeans IDE. Clever-cloud as a storage platform is used with FS Bucket to create different servers. Three servers created for testing purpose.

IV. CONCLUSION

In this paper, the challenges of privacy preserving in cloud storage are discussed. To conquer these challenges the ORAM algorithm is applied to achieve privacy-preserving access to big data in clouds. The load unbalance phenomenon have occured after deploying ORAM-based storage to multiple servers. To overcome this problem the ILB algorithm is used which solves data placement problem to achieve load balance.

V. REFERENCES

[1] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel, "Finding a needle in haystack: Facebook's photo storage," in USENIX OSDI, 2010, pp. 1–8.

[2] M. Islam, M. Kuzu, and M. Kantarcioglu, "Access pattern disclosure on searchable encryption: Ramification, attack and mitigation," in Network and Distributed System Security Symposium, 2012.

[3] Zhang, F., Q. M. Malluhi, T. Elsayed, S. U. Khan, K. Li, and A. Y. Zomaya.. "CloudFlow: A Data-Aware Programming Model for Cloud Workflow Applications on Modern HPC Systems." Future Generation Computer Systems 51: 98–110, 2015.

[4] E. Stefanov, M. van Dijk, A. Juels, and A. Oprea, "Iris: A scalable cloud file system with efficient integrity checks," in Proceedings of the 28th Annual Computer Security Applications Conference, 2012, pp. 229–238.

[5] K. D. Bowers, A. Juels, and A. Oprea, "Hail: A high-availability and integrity layer for cloud storage," in Proceedings of the 16th ACM Conference on Computer and Communications Security, 2009, pp. 187–198.

[6] C.Z. Xu and F.C.M. Lau. Optimal parameters for load balancing with the diffusion method in mesh–networks. Parallel Processing Letters, 4(2):139–147, 1994.

[7] O. Goldreich and R. Ostrovsky, "Software protection and simulation on oblivious rams," Journal of the ACM (JACM), vol. 43, no. 3, pp. 431–473, 1996.

[8] Zhan, Z. H., X. F. Liu, Y. J. Gong, J. Zhang, H. S. H. Chung, and Y. Li. "Cloud Computing Resource Scheduling And a Survey of its Evolutionary Approaches." ACM Computing Surveys (CSUR) 63: 1–33, 2015.

[9] P. Williams and R. Sion, "Single round access privacy on outsourced storage," in ACM CCS, 2012, pp. 293–304.

Cite this article as :

P. B. Niranjane, "Load Balancing Using Oblivious RAM for Privacy-Preserving Access to Big Data in Cloud ", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 2, pp. 885-889, March-April 2023. Available at doi : https://doi.org/10.32628/IJSRST523102140

Journal URL : https://ijsrst.com/IJSRST523102140