

A Survey on an Optimum Load Balancing for Large Data on Cloud Using Machine Learning

Shailesh B Galande¹, Sachin Patil²

¹GHRCEM, Wagholi, Pune, Maharashtra, India

²Assistant Professor, GHRCEM, Wagholi, Pune, Maharashtra, India

Abstract: Bulky Data denotes to the huge volumes of data gathered since ages which is tedious to examine and handle using common database management tool. So we have to implement large data with data mining using linear programming or machine learning. However user can perform operation at dynamically on cloud computing. These concepts not only implement data mining operation on cloud but also perform load balancing operation. With cloud data services, it is common place for data to be not only stored in the cloud, but also to be shared across multiple users. This system of Cloud storage enables loading of data in the cloud servers efficiently and allows the user to tackle with the data without any difficulty of the resources. In the previous system, the data is deposited in the cloud using active data operation with a single cloud service provider. The cost and the Quality of service provided to the user are limited as provided by CSP. In this paper, the partitioning method is anticipated for the data repository which stores data on multiple CSPs, ensuring data security by partitioning the data which avoids the local copy at the client side. This method provides and guarantees high cloud repository integrity, improved error localization and easy identification of misbehaving server. To attain this, mobile data integrity checking concept and partition allocation using linear programming is implemented to improve the performance of cloud storage.

Keywords: Bulky Data, Machine Learning, CSPs, QoS, Cloud computing, Load balancing

1. Introduction

The technology used in Cloud storage is indeed the most well-known and innovatory technology in this phase of hi-tech spreads. It consists basically of repository where digitized data is stored. Multiples servers geographical positioned at distant locations make up for the physical environment of the system. There are servers which are governed by a presenting company. These Cloud Storage Providers (CSPs) are in authority for holding as well as providing the data obtainable and accessible whenever required and apart from that they are also in authority for keeping the physical setup active and dynamic. The technology is so well known that we barely comprehend its importance that all of us access easily our mails, accounts on societal sites as well as other things on the move but without any worrying effort. Such things are possible because of the cloud storage technology. Cloud computing is a solution to the tedious work of carrying and handling our personal system. Cloud enables us to carry data when mobile in spite of any risks.

A cloud storage system facilitates depositing of data in the cloud server proficiently and effortlessly allowing user to manage resources. Cloud storage is classified into basically two categories viz. private, public or hybrid (Fig 1). Private cloud storage is limited to a specific association and data security risks are less as compared to public cloud storage. In the systems that are today data is stored in the cloud using active data operation, but the problem lies in the making of an additional copy on user side for additional apprising and verification of the data loss. Partitioning method is suggested for the data storage which avoids the creation of local copy at user side. The extraordinary addition of cloud computing, built on the fixed research

fields of distributed computing, web services, networks, utility computing and virtualization.

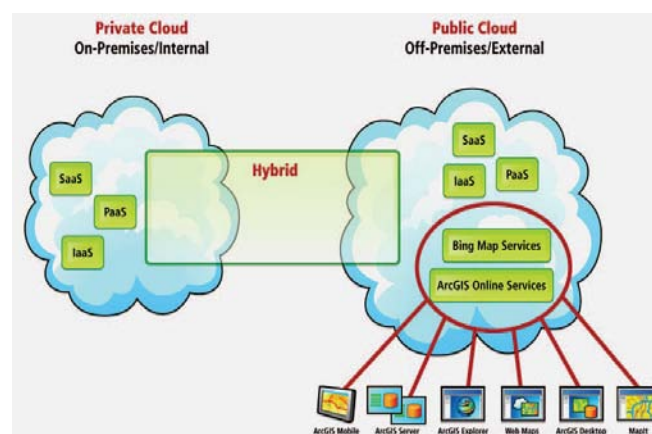


Figure 1: Cloud Services Architecture

Diverse techniques have been integrated to make certain high cloud storage integrity, upgraded error localization and easy identification of defiant server. Cloud storage enactment is improved using integrity scrutiny concept. The emphasis is centered on storing data in minimum space in less time and less computing cost. Basically, the cloud is categorized in three main components viz. Clients, Data Centers and Distributed Servers. (Fig 2)

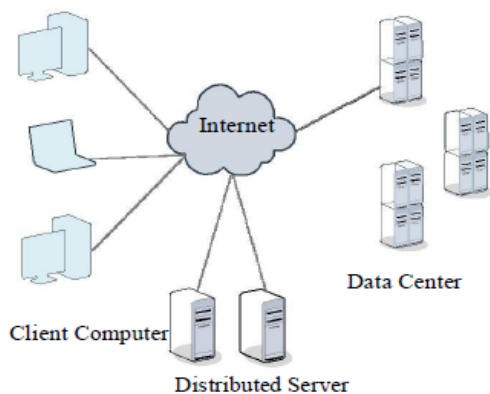


Figure 2: Cloud Computing Components

2. Related Work

[1] Cloud storage is the broader theme but the system [1] mainly focuses on private cloud storage. A private cloud storage is built using commodity machines. As the storage space is limited, judicious use is a compulsion. Techniques like duplication have been used to create unique instances of data in the storage. Imbalance is caused due to expansion of cluster which is overcome by using centralized load balancing technique.

[2] To remove the overhead of query retrieval time the system [2] comes with the concept of look up tables as a

way to provide fine grained partitioning for transactional database applications. Using look up tables query result time is greatly reduced. Twitter, TPC-E and Wikipedia were used for partitioning using look up tables. On the above mentioned applications from 40 percent to 300 percent performance hike was recorded.

[3] Symmetric key algorithms for example Advanced Encryption Standards (AES) are used in PDP scheme proposed in [3] Data availability i.e. making sure that data is available any time it is accessed and data error recovery are not given much importance in the system.

[4] The paper [4] emphasizes a need on data security and data privacy taking the storage of pictures by users in cloud into picture. As digital images may include sensitive information they need to secure and privacy be maintained. Face recognition and stripping algorithms have been proposed in which sensitive data remains with the Cloud user. There is a problem with the system implementation with true negatives, i.e. the data is classified as non-sensitive but it has to be classified as sensitive. So this needs a further look into it.

3. Existing System

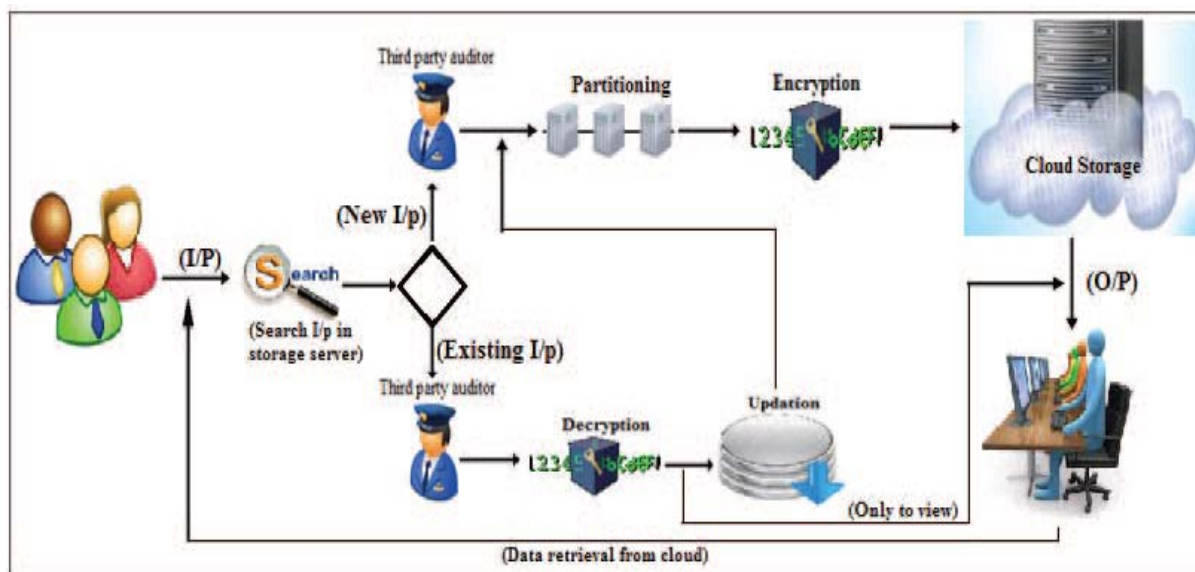


Figure 3: Partitioning and Domain Integrity Checking For Data Storage in Cloud Computing

The existing system represents how data is stored and distributed on the cloud using partitioning to provide security and availability across two servers.

3.1 Work Flow

3.1.1. Client Side

- a) File Selection: The user/users select a File to upload on the server.
- b) Sending File: The client/clients sends respective file to Third Party Auditor (TPA) across the network using a File Transfer Protocol (FTP).

3.1.2. TPA

TPA performs six types of steps on the received file/files from the respective clients for storage across cloud with security and availability.

- a) Receiving File
Third Party Auditor Receives File from the client/clients.
- b) Partitioning File
TPA partitions file received from the client/clients.
- c) Digital Signature Extraction
TPA extracts Digital Signature of each file partition.
- d) Secret Key Generation

After partitioning, Third Party Auditor generates Secret keys for each partition respectively.

e) Encryption

TPA encrypts each partition using respective secret keys.

f) Storing Partition Sequence

TPA stores Partition Sequence, Signature, Keys and File attributes on its own server.

3.1.3 Server Side

a) Sending Partitions

Third Party Auditor sends the respective partition to the respective storage.

The Respective Storage Server receives the respective File from the Third Party Auditor.

b) Storing

The storage server stores the partition received from the TPA.

4. Proposed System

The end user fires a query when he needs a certain piece of information simply using his browser. The file is send over the network where it is received by Third Party Auditor (TPA), it receives the file and partitions it into blocks and generates keys for each partition and encrypts each partition using the respective keys. TPA stores partition sequence, signature, keys and file attributes on its own server. Naive Bayes Classifiers are used and the file is forwarded to the database servers where it is stored in the partitioned form for quick access. When a user fires a query for retrieval of data the partitioned file stored on the database are merged together again to form the original file.

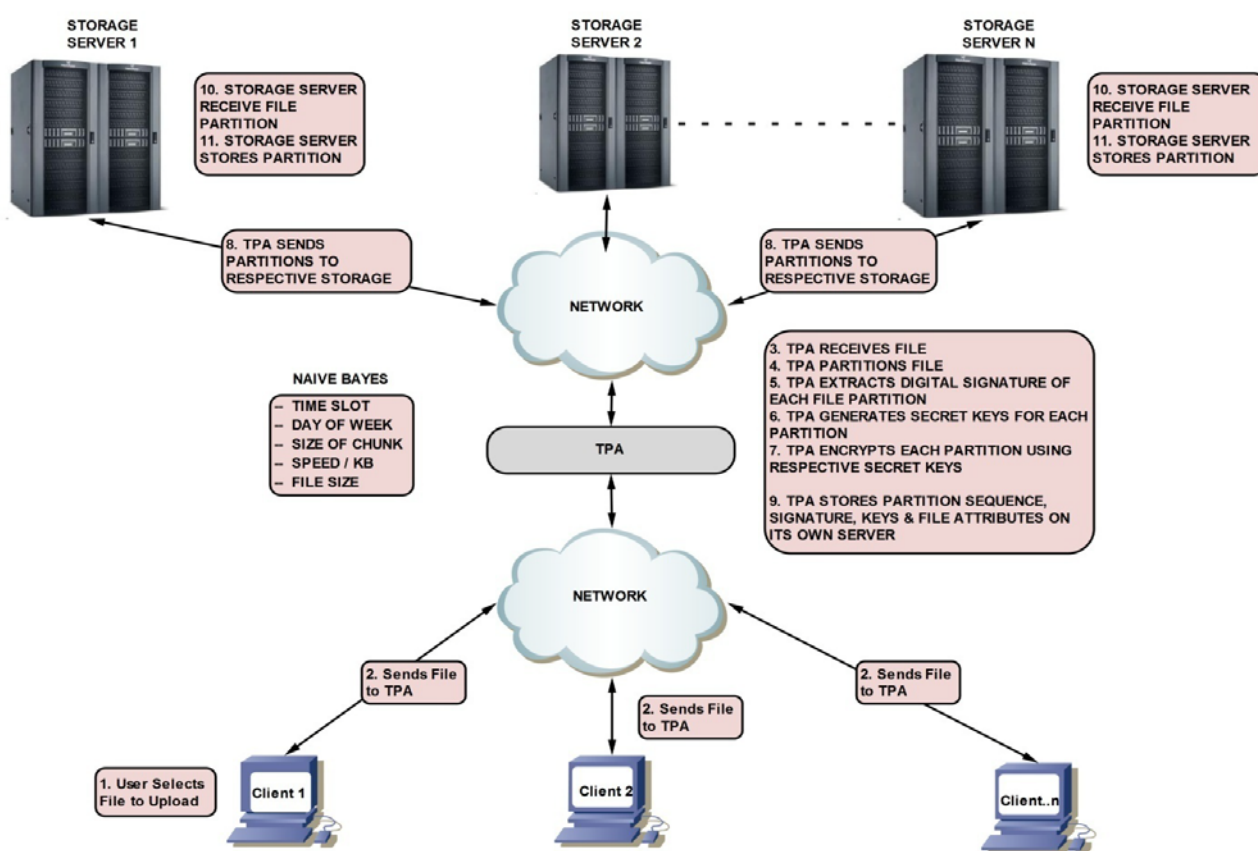


Figure 4: Partitioning and Domain Integrity Checking for Data Storage in Cloud Computing

The diagram shown above shows the proposed system. The system consists of end users, a Third party auditor, a network and the multiple servers for storing the data.

- USER: The end user is the actual user which queries the database in the servers using browser installed in his system.
- TPA: Basically is meant for auditing in the system. It uses various algorithms for encryption, decryption and authentication purposes.
- Network: The network used is Internet.
- NBC: Naive Bayes Classifier
- Servers: The servers store data in the partitioned form. After a query is fired the data is organized and sent to user

5. Limitations

Handling large or bulky data is a tough job; to explore large data analysis has to be done at data, model, and system levels. High computing platforms are also required which is a major constraint. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. The TPA also poses certain limitations to the algorithms used to encrypt or decrypt the data which is prone to various types of attacks; complexity of the system is also on the higher side which makes it difficult to comprehend.

6. Conclusion

The Proposed system is efficient system considering the security and other aspects of the cloud storage technology. Partitioning of data makes data access easy and quick. Data stored is highly secured using the cryptography algorithms and digital signatures. It integrates some new concepts like data security, storage optimality, file integrity and authentication access which are not present in the current system. NBC and HACE theorem make the system much more robust as compared to existing ones.

References

- [1] Prabavathy.B, Priya K, Chitra Babu, "A Load Balancing Algorithm for Private Cloud Storage"
- [2] Aubrey L. Tatarowicz, Carlo Curino, Evan P. C. Jones, Sam Madden, "Lookup Tables: Fine-Grained Partitioning for Distributed Databases", Massachusetts Institute of Technology, USA.
- [3] C. Wang, K. Ren, W. Lou, and J. Li "Towards Publicly Auditable Secure Cloud Data Storage Services", IEEE Network Magazine, vol. 24, no. 4, pp. 19-24, July/Aug.2010.
- [4] Rene Leistikow and Djamshid Tavangarian "Secure Picture Data Partitioning for Cloud Computing Services", 27th International Conference on Advanced Information Networking and Applications Workshops 2013.
- [5] Yanhong Zhai and Bing Liu "Web Data Extraction Based on Partial Tree Alignment".
- [6] Yunhua Deng and Rynson W.H. Lau, senior member IEEE, "On Delay adjustment for Dynamic load balancing in distributed virtual environment", April 2012.
- [7] Bin Dong, Xiuqiao Li, Qimeng Wu, Limin Xiao and Li Ruan, "A dynamic and adaptive load balancing strategy for parallel file system with large-scale I/O servers," Journal of Parallel Distribution. Computer.72, pp. 1254–1268, 2012.
- [8] Pragati Priyadarshinee and PragyaJain, "Load Balancing and Parallelism in Cloud Computing," International Journal of Engineering and Advanced Technology, Vol. 1, Issue. 5, 2012, pp. 486-489.
- [9] Bhushan Lal Sahu, Rajesh Tiwari, "A comprehensive study on Cloud computing", International Journal of Advanced Research in Computer Science and Software Engineering, volume 2, issue 9, September 2012, ISSN: 2277 128X.
- [10] Ratan Mishra, Anantjaiswal, Ant colony optimization: A Solution of load balancing in cloud, International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012.
- [11] D. Escalante, Andrew J. Korty, "Cloud Services: Policy and Assessment", Educause review July/August 2011.