Survey of Data Deduplication By Using Cloud Computing

Bhairavi Kesalkar[1], Dipali Bagade[2], ManjushaBarsagade[3], Namita Jakulwar[4]
Prof. Shrikant Zade[5]

[1][2][3][4]Scholar (UG),CSE Department, PIET, Nagpur, Maharashtra, India
[5]Assistant Professor, CSE Department, PIET, Nagpur, Maharashtra, India

Abstract:

Cloud computing has quickly become one of the most significant field due to its evolutionary services provided model of computing not only in the IT industry but also in the software and hardware industry. This mechanism came up with increasing flexibility, scalability and reliability; while decreasing the operational and support cost. Due to the cloud computing, it becomes easy for managing the stuffs related as well as provides many features which cannot be replaced by anyone. It is a way difficult as well as effective in its own. Providing security is a major concern as the cloud data are stored and accessed in a remote server with the help of by the cloud service provider .Translation of efficient storage and security for all data is very important for cloud computing. Securing and privacy preserving of data is of high priority when it comes to cloud storage. Therefore to provide efficient storage for cloud data owners and provide high security for data this paper proposes Cloud Computing. Intrusion , detection and prevention are performed manually by network operators in the existing system. Data deduplication technique allows the cloud users to manage their cloud storage space effectively by avoiding storage of repeated data's and save bandwidth. The data are finally stored in cloud server. To ensure data confidentiality the data are stored using encryption.

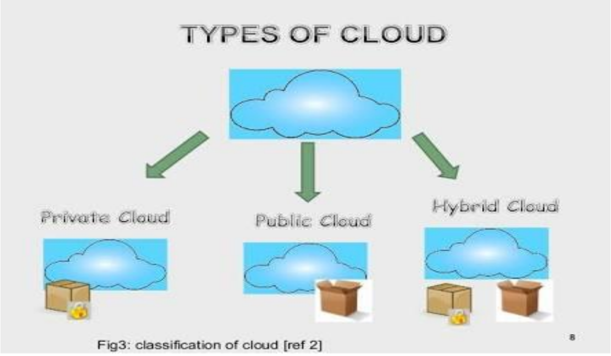
Keywords: De-duplication, Cloud computing.

ISSN : 2348-5612 © URR
9 770234 856124

1. INTRODUCTION

1.1Cloud computation

Every industry either the smaller ones or the large ones, all of them have the large amount of data that they need to manages, want it to secure and have privacy means they do not want their data to get accessed by any other third party or unauthorized user. Previously for this purpose, these industry store that data in their personal computers or small servers. But by the time their data became unmanageable by them due to less space as well as time and people and they felt difficult in managing their data. Then the concept of cloud was introduced. Some of the organizations who have got more space means server started giving their server on rent for their profit. They started giving their space as well as they started providing the facility or services of managing their data for those organizations or individual. As per the requirement of the companies there are four types of cloud which got introduced. These cloud types are actually known as deployment model of cloud are 1) Private Cloud 2) Public Cloud 3) Hybrid Cloud.



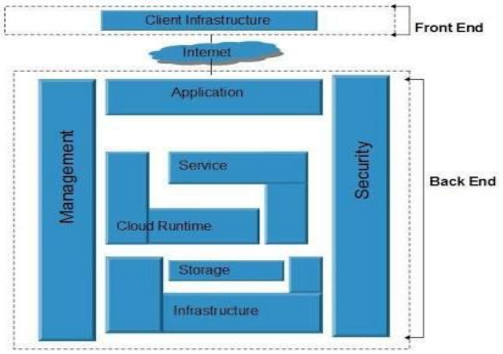
Fig(b):Types of Cloud

1)Public Cloud: It is the most known cloud in all because it is easily usable any anyone. It is kind of straight forward cloud computing. Public cloud is accessible by more than one organization.

2)Private Cloud: It has some similar features to public cloud like scalability and self-service. The private cloud is generally owned by the large organization because it costs very much for accessing and managing. Private cloud is accessed by single organization

3)Hybrid Cloud: It has the infrastructure which is made of the combination of the segments of public cloud which is the third party and the private cloud that is the client side. It has the security and it is flexible.

1.2 Architecture of Cloud:



Fig(a). Architecture of cloud computation

Cloud computing is divided into two part first one is the front end which is shown in the user side, and second is the back end which is the "cloud" .Both the are connected to each other through INTERNET. Client's computer and the application request to access the cloud computing system are include in front end .The back end of the system are the various computer server and data storage system that create the "cloud" of computing system. A third party control the flow of both ends , monitoring traffic and client demand called as a central server administrators. A special kind of software use to flow a set of rules and protocols called as the middleware which is cloud operating system.

2. Data De-duplication

Data deduplication is rapidly growing technique now days especially in backup storage due to reduction in cost of storage. Data deduplication is very important in management of data because it will store only unique data among duplicate data copies. Data Deduplication is efficient technique to handle these large duplicate data. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy [1].Data deduplication can be source based and target based. In the target based Data Deduplication user will upload their data and deduplication will take place at target side. So target based approach can improve storage utilization but cannot save bandwidth as whole data needs to be transferred at target side in source base deduplication clients will check at storage side whether the data copy is already exists or not, that means deduplication will be perform at client side and then after only unique copy will be stored. So source based approach can improve bandwidth as well as storage .There is also granularity based deduplication: File level deduplication . In file level only unique copy of file will be stored and duplicate copy will be discard .

The deduplication gives noteworthy benefits but security and data confidentiality is still sensitive issues. So, usual way to provide security is encryption. But there is confliction between data deduplication and encryption. One more thing needs to be there in data deduplication is authorized deduplication in which users would have set of privileges because in many of applications differential authorized duplication is needed. User can not check duplicate out of his privilege set.

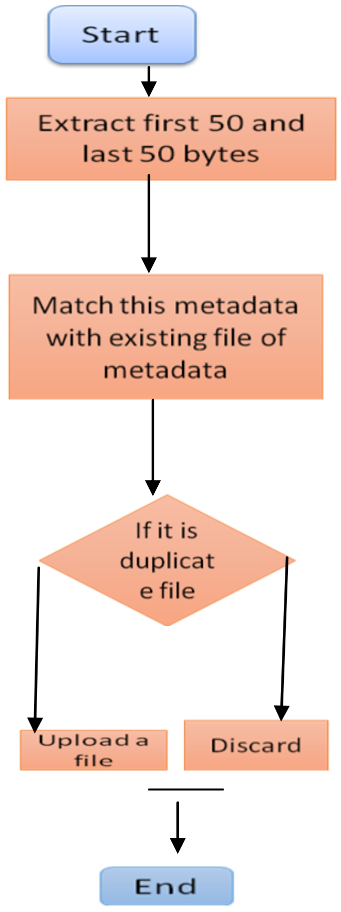

Start
Extract first 50 and last 50 bytes
Match this metadata with existing file of metadata
If it is duplicat e file
Upload a file
Discard
End

Fig 1: Flow chart of data- deduplication

3. Existing System

The cloud environment is a large open distributed system. It is important to preserve the data, as well as privacy of the users. Existing techniques in cloud storage are ***Redundant Array of Cloud Storage (RACS):*** This RACS is used to stored data over multiple vendors. This is act as a proxy and performs the operation between the client and server. This method is simple and easy to work with. The drawback is single proxy cloud easily become bottleneck. ***Secure Overlay Service (SOS):*** This SOS is mainly used to solve the distributed denial of service, the idea provide by this solution is very complicated. It is unclear about the optimal solution. ***Vanish****:* This technique of storage is all data become unreliable after some specific period of time for the security purpose. The data will be deleted after the particular time period with the knowledge of the owner who created the data. This technique is more expensive and it requires large Distributed Hash Table (DHT). ***FADE:*** This FADE is secure overlay in cloud storage with assured deletion. The data owner can be sure of the deleted file. Only the deleted part of the file is considered not the accessing data. This method is more complex to implement.

Limitation/Drawbacks of existing system:

1) Existing system contains Manual work.

2) Data deduplication does not work with traditional encryption technique while using data deduplication technique is should not reduce fault tolerance mechanism.

3. Proposed Work

Our proposed technique is deduplication scheme. This technique eliminates the content of same of data and stored once. If the cloud service providers implements this technique they can reduce the storage space and the uploading bandwidth. This method reduces the cost of the storage and increase the security of the data.

Attacks possible in deduplication: ***Predicting File****:* In this attack suppose the attacker wants to find out whether user1 possesses a file, File A. He will upload a copy of file A if the file gets uploaded this will indicate that the file is not possessed by the user1 otherwise the attacker easily finds out the file posses by the user.

Creating secret channel: In this attack the attacker install malicious software on the users A machine. This malicious software creates the side channel between the user and the attacker. Then the attacker easily finds out the users information. ***The content distribution attack:*** In this attack the attacker distribute the some file in the storage system to check whether hash value is matched with the user A file, if it is matched then that file can be easily compromised.

Solution for the attacks: Solution for those attacks is given by encrypt the file before uploading, performing target based deduplication. Randomization, Gateway based.

4. Literature Survey

From this paper [1] they conclude that Cloud Computing trend is rapidly increasing that has an technology connection with Grid Computing, Utility Computing, Distributed Computing. Cloud service providers such as Amazon IBM, Google's Application, Microsoft Azure etc., provide the users in developing applications in cloud environment and to access them from anywhere. Cloud data are stored and accessed in a remote server with the help of services provided by cloud service providers. Providing security is a major concern as the data is transmitted to the remote server over a channel (internet). [2]

From this paper [2] they conclude that shows the framework for designing the trusted platform for the cloud computing system. Data stored in third party storage systems like the cloud might not be secure since confidentiality and integrity of data are not guaranteed. Though cloud computing provides cost-effective storage services,. In this thesis, a solution to the problem of securely storing the client's data by maintaining the confidentiality and integrity of the data within the cloud is developed. Five protocols are developed which ensure that the client's data is stored only on trusted storage servers, replicated only on trusted storage servers, and guarantee that the data owners and other privileged users of that data access the data securely.[4].

From this paper [3] they conclude that Cloud computing is an architecture for providing computing service via the internet on demand and pay per use access to a pool of shared resources namely networks, storage, servers, services and applications, without physically acquiring them. So it saves managing cost and time for organizations. Many industries, such as banking, healthcare and education are moving towards the cloud due to the efficiency of services provided by the pay-per-use pattern based on the resources such as processing power used, transactions carried out, bandwidth consumed, data transferred, or storage space occupied etc.. There are various research challenges also there for adopting cloud computing such as well managed service level agreement (SLA), privacy, interoperability and reliability. This research paper also analyzes the key research and challenges that presents in cloud computing and offers best practices to service providers as well as enterprises hoping to leverage cloud service to improve their bottom line in this severe economic climate.[8]

From this paper [5] they conclude that Cloud Computing has been envisioned as the next generation architecture of IT Enterprise. In contrast to traditional 3 solutions, where the IT services are under proper physical, logical and personnel controls, Cloud Computing moves the application software and databases to the large data centers, where the management of the data and services may not be fully trustworthy. This unique attribute, however, poses many new security challenges which have not been well understood. In this article, we focus on cloud data storage security, which has always been an important aspect of quality of service. To ensure the correctness of users' data in the cloud, we propose an effective and flexible distributed scheme with two salient features, opposing to its predecessors. By utilizing the homomorphic token with distributed verification of erasure-coded data, our scheme achieve the integration of storage correctness insurance and data error localization, i.e., the identification of misbehaving server(s). Unlike most prior works, the new scheme further supports secure and efficient dynamic operations on data blocks, including: data update, delete and append. Extensive security and performance analysis shows that the proposed scheme is highly efficient and resilient against Byzantine failure, malicious data modification attack, and even server colluding from this paper they conclude that attacks.

5. Conclusion

Thus this paper compresses the data by removing the duplicate copies of identical data and it is extensively used in cloud storage to save bandwidth and minimize the storage space. To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data before outsourcing. For better data protection, this paper talks about the issue of data deduplication authorization.

6. References

[1]Jin Li ,Yan Kit Li ,XiaofengChen ,Patrick P.C. Lee and WenjingLou ”A Hybrid Cloud Approach for Secure Authorized Deduplication” IEEE Transactions On Parallel And Distributed System Vol.26,No.5, May2015

[2]Vishwakarma R Ganesh “R. VelumadhavaRao, K. Selvamani, “Data Security Challenges and Its Solutions in Cloud Computing” @ICCC 2015.

[3]Android College Management System”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 4, April 2016, ISSN: 2278 – 1323

[4]Rohit Bajaj, “Developing framework for secure storage in cloud computing system” @IJNIET, Vol 1 issue 3, February 2013[5] S .R .Bharamagoudar , Geeta R.B., S .G .Totad “Web Based Student Information Management System”, International Journal of Advanced Research in Computer and Communication Engineering -June 2013, ISSN : 2319-5940.

[6] P. Anderson and L. Zhang, ‘‘Fast and Secure Laptop Backups with Encrypted De-Duplication,’’ in Proc. *USENIX LISA, 2010, pp. 1-8.*

[7] J.R. Douceur, A. Adya, W.J. Bolosky, D. Simon, and M. Theimer, ‘‘Reclaiming Space from Duplicate Files in a Serverless Distributed File System,’’ in Proc. *ICDCS, 2002, pp. 617-624.*

[8] Rabi Prasad Padhy, ManasRanjanPatra, Suresh Chandra Satapathy, “Cloud Computing: Security Issues and Research Challenges” @IRACST 2011

[9]Q.Wang, C.Wang, J. Li, K. Ren, and W. Lou. Enabling public verifiability and data dynamics for storage security in cloud computing. In European Symposium on Research in Computer Security (ESORICS '09), volume 5789 of Lecture Notes in Computer Science, pages 355{370. Springer, 2009

[10] JinMei-shan, QiuChang-li,LiJing, "The Designment of Student Information Management System Based on B/S Architecture", IEEE Computer Society ,978-