International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

Fast and Secure Cloud Big Data Dedup.

Ms. Pallavi Maladkar[1], Vaishnavi R Bhadgaonkar[2], Namrata S Jamdar[3], Shraddha R Gadekar[4], Mukta U Awate[5]

[1]Lecturer, Computer Engineering, Pimpri Chinchwad Polytechnic, Nigdi, Pune, India.
[2,3,4,5]Third Year Diploma Student in Computer Engineering, Pimpri Chinchwad Polytechnic, Nigdi, Pune, India.

Abstract – *Previously, data was stored on traditional hard disks, servers, etc. as the data was limited. But data generated today is in terms of petabytes and number of people or users or customers accessing this data is also increasing day by day. Hence, the major problem of this Big Data is not related to storage, but is related to efficient space balancing and handling. So, the main goal of our project is to increase maximum utilization of the memory. Thus, main idea behind our project is preventing data duplication in companies, industries, etc. So, our project provides secure deduplication method for efficient storage management. It provides Encryption of data, Random Key Generation for verification and Email system for providing the activation key to user. It also consists of variable length i.e. block level deduplication and Hash function for comparing blocks and is very easy to use. Our project also provides verify option to check where the file is safe or not.*

***Key Words*: Data Deduplication, Block-level Deduplication, Encryption, Random Key Generation, Hash Function.**

1. INTRODUCTION

It is called cloud computing because the information being accessed is found in "the cloud" and does not require a user to be in a specific place to gain access to it. This type of system allows employees to work remotely. Companies providing cloud services enable users to store files and applications on remote servers, and then access all the data via the internet. It is a delivery of on-demand computing resources - everything from applications to data centers - over the internet on a pay-for-use basis. It provides scalability of resources, only pay for what you use and needed services are provided by self-service accesses. It has three main types based on services:

Software as a Service (SaaS): SaaS involves the licensure of a software application to customers. Licenses are typically provided through a pay-as-you-go model or on-demand. This rapidly growing market could provide an excellent investment opportunity

Infrastructure as a Service (IaaS): Infrastructure as a service involves a method for delivering everything from operating systems to servers and storage through IP-based connectivity as part of an on-demand service. Clients can avoid the need to purchase software or servers, and instead procure these resources in an outsourced, on-demand service.

Platform as a Service (PaaS): Of the three layers of cloud-based computing, PaaS is considered the most complex. PaaS shares some similarities with SaaS, the primary difference being that instead of delivering software online, it is actually a platform for creating software that is delivered via the internet.

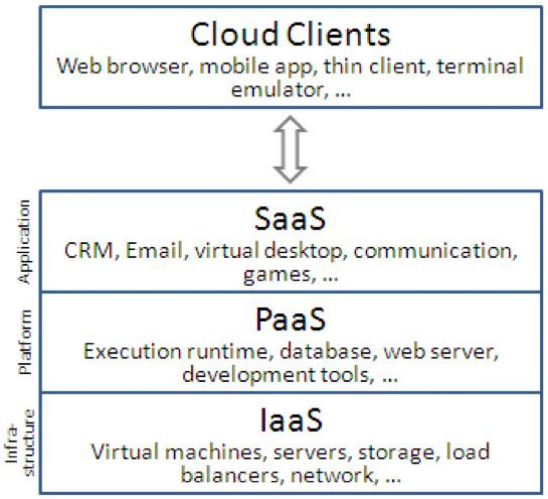

Fig 1.1. Types of Cloud Based on Service

Also it has three main types based on Location:

Public: A public cloud is one based on the standard cloud computing model, in which a service provider makes resources, such as virtual machines (VMs), applications or storage, available to the general public over the internet. Public cloud services may be free or offered on a pay-per-usage model.

Private: Private cloud means using a cloud infrastructure (network) solely by one customer/organization. It is not shared with others, yet it is remotely located. If the cloud is externally hosted. The companies have an option of choosing an on-premise private cloud as well, which is more expensive, but they do have a physical control over the infrastructure.

Hybrid: Hybrid cloud, of course, means, using both private and public clouds, depending on their purpose.

For example, public cloud can be used to interact with customers, while keeping their data secured through a private cloud.

Community Cloud: Community cloud implies an infrastructure that is shared between organizations, usually with the shared data and data management concerns. For example, a community cloud can belong to a government of a single country. Community clouds can be located both on and off the premises.

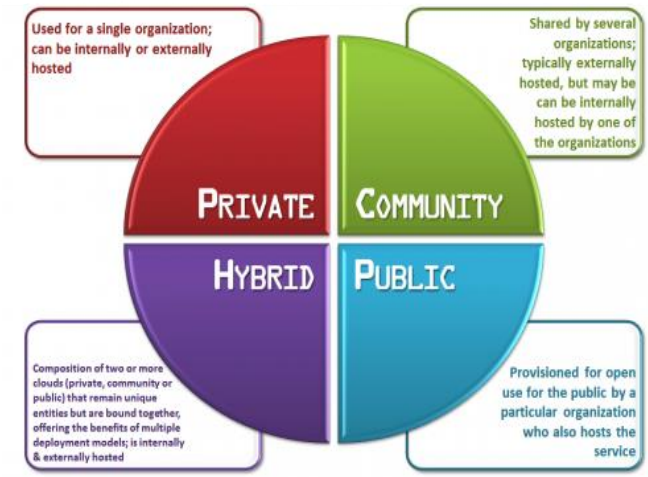

Fig 1.2.Types of Cloud based on Location

As Cloud technology is advancing every day, new challenges are encountered. Now the space and network charges are not a problem. The big problem in cloud computing is space balancing and handling. So these Cloud storage providers heavily rely on Data Deduplication to save storage costs by only storing one copy of each uploaded file.

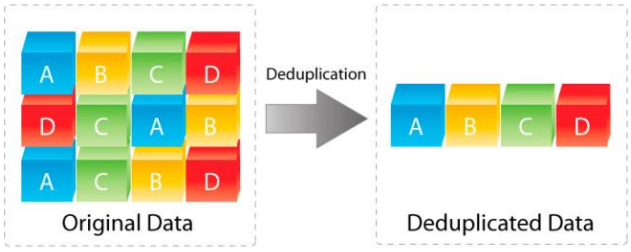

Fig 1.3. Deduplication.

Deduplication identifies and locates the duplicate data. It then eliminates duplicate copies of repeating data and saves the space for data that needs to be physically stored. The existing system and it is done by two ways : a)File-level deduplication and b)Block-level deduplication. The existing traditional systems of Deduplication was not secured i.e. the privacy of data holders as encrypted data introduced new challenges for cloud data deduplication i.e. tedious to process and was not divided into chunks for more efficient Deduplication. The copy uploaded by a specific user was not available for other users causing redundancy of data individually. The proposed system has provided block level Deduplication i.e. Fixed size or variable size of blocks are created for efficient storage of files and Proof of Ownership is achieved. Also it is a Secure Deduplication - data encryption is provided for ensuring the security of the data holders. Also Electronic-mail is used as a method of communication.

2. LITERATURE SURVEY

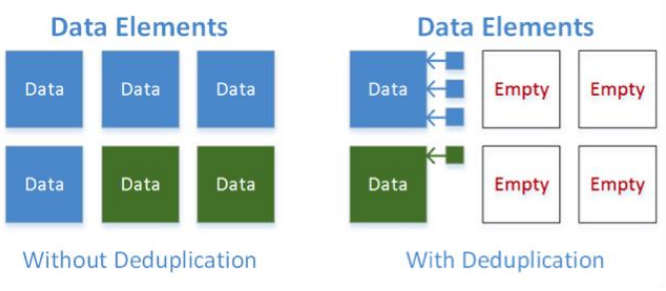

Fig 2.1 Data without and with Deduplication Technique.

The most surfaced issues were as follows:

a)Large amount of data was being stored on Clouds ,

b)The stored data could be hacked easily i.e. there was lack of Security (Encryption),

c)Efficient deduplication was not taking place i.e. file names or complete file contents were checked. d)Integrity and confidentiality of data could be compromised easily.

e)It was less reliable.

f)User Verification was not accurate.

3. RELATED WORK:

IEEE papers regarding this topic are:

Improving Efficient Reliability Based Secure Deduplication Technique on Encrypted Big Data: In this paper Security and Confidentiality are given priority. Sometimes, because hacking attackers cloud is unsafe. Therefore authorized de-duplication will only give effective result. To maintain data in high priority without hacking on cloud in de-duplication is must. Reduction and saving energy is securely maintained for authorized de-duplication in cloud. For reducing attacker we proposed the CIDS algorithm which means network based attacker, Masquerade attackers and host based attackers. IDS handle easy targets for cloud user. CIDS architecture is scalable and elastic without central coordinator. The cloud users each one has secret key for file sharing through that we can save the file. While uploading file we can access the data is duplicate or not duplicate also know as it is which type of duplicate which means file level deduplicate and block level deduplicate. Cloud service provider can find the file is which type of duplicate file. If it is duplicate it's automatically move to deactivation then if it is non-duplicate file means activate the file automatically.

Transparent Data Deduplication in the Cloud: In this paper, we propose a novel storage solution, ClearBox, which

allows a storage service provider to transparently attest to its customers the deduplication patterns of the (encrypted) data that it is storing. By doing so, ClearBox enables cloud users to verify the effective storage space that their data is occupying in the cloud, and consequently to check whether they qualify for benefits such as price reductions, etc. ClearBox is secure against malicious users and a rational storage provider, and ensures that files can only be accessed by their legitimate owners.

Cryptographic Tuning to Minimize Storage Requirement on Cloud Using De-Duplication Mechanism: The notion of authorized data de-duplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented new de-duplication check to supporting authorized duplicate-check tokens of files that are generated by the cloud server with private keys. The proposed system is secure from insider and outsider attacks. The proposed system implement new prototype model in which we used convergent encryption with modification version to deal with brute force attack using Cryptographic tuning to make better authorized de-duplication technique, etc.

Then tutorial on Cloud Computing Tutorials from tutorialspoint.com, edureka, etc. Seminars on Cloud Computing and Applications from Department of Computer Science and Engineering Indian Institute of Technology, Bombay.

4. OVERVIEW OF PROPOSED SYSTEM:

4.1 PROBLEM STATEMENT:

Problems regarding user friendliness, authentication, etc. were faced. Also, main focus is on balancing space so as to increase utilization.

4.2 SOLUTION:

In this project, we propose a system that is highly reliable and secure. Here, the first step is the file being stored will be encrypted by a key formed by Random Key Generation Algorithm (i.e. it will generate the key from the contents of the file only and the user doesn't have to memorize or share the key with receiver i.e. Dekey) and then it will be checked in the Cloud, if the file is unique then it will get stored but if the file is already present into the cloud then that file will not be stored and Proof Of Ownership (i.e. the ownership pointer of the document) will be given to that second user also.

But, if the file stored has a different name from those stored on the cloud, then further it will be divided into chunks i.e. blocks and using these blocks we will generate the Hash Tag by using Convergent Encryption to check within the cloud.

Will retrieving the file from the cloud it will be decrypted first by the key generated by Random Key Generation Algorithm which is stored and then will be delivered to the User.

4.3 ARCHITECTURE OF THE SYSTEM:

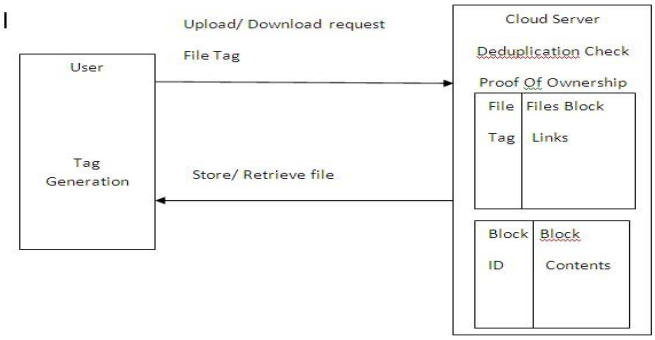

Fig 4.1. Cloud Data Deduplication Diagram

Here we can see how actually the data is being deduplicated. User A will upload a file named "ABC.txt" with contents "Hello" and according to the contents of file the key will be generated and it will be encrypted and will be stored by the Hash Tag's value. When User B uploads the file named "ABC.txt", it will be checked whether it is already existing or not, if it is already existing then the file will not get stored and Ownership pointer will be linked to User B. But, if User B enters a file named "PQR.txt", then as the file names are different, PQR.txt must get stored, but as we are proposing efficient storage, hence the contents of files "ABC.txt" (is "Hello")and "PQR.txt" (assuming "Hello Hi") will get divided into blocks and then the blocks will get checked. User A and User B both are having Hello blocks hence that will be stored as a single block and Hi block from User B will be stored as different unique block.

There are four main modules in our Project:

- USER WITH EMAIL
- TPA
- CLOUD
- BACKUP SERVERS

4.3.1 THE USER:

- This is the first module which consists of two main parts, namely the User Login and if the user is new, User Registration.
- The user can upload a file as well as download an file uploaded on Cloud.
- Whenever a new user logs in, an Activation Key is generated and this key is provided to the authorized users Gmail account making it more secured and user friendly.
- Due to email facility, User confirmation or validation of user takes place making it more secure. The tedious work of collecting key from the administrator can be overcomed.

- Here the main part of division of data into blocks and hash key generation takes place. Hash Key Function is used as an identification number to each block of data.

- Also, data divided into the block gets encoded and is ready to be stored on the cloud. User send this upload request to TPA.

4.3.2 THE TPA(Third Party Auditor):

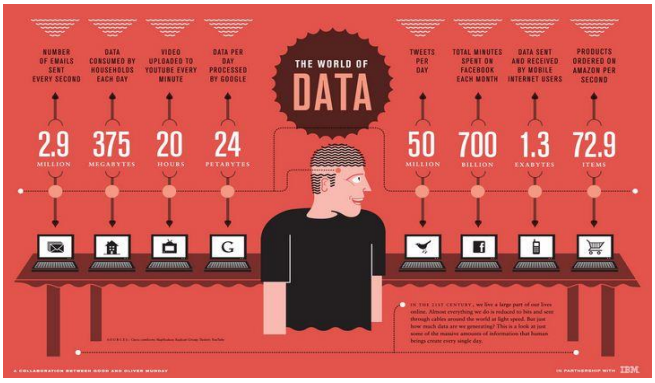

Fig 4.2. Data Uploaded per day.

- The TPA i.e. Third Party Auditor acts an administrator, but its main job is to control the number of requests on the Cloud.

- While managing the requests he sorts the request based on the priorities of each request.

- If the TPA is absent in an organization and due to tremendous requests for data as shown in fig 4.2 simultaneously on cloud, the chances of data loss and slow transfer rate are high.

- So, the TPA module is used. Then TPA forwards the request to Cloud.

4.3.3 THE CLOUD:

- Here, in this module, data is verified first i.e. it is checked where it is hacked or not and is stored.

- It is made available for Users anytime they need it as an response to their requests.

- Whenever a copy in block format is stored on Cloud, it is also stored in our Backup Servers.

4.3.4 THE BACKUP SERVERS:

- The Backup servers are used for avoiding data loss. If any failure occurs the data can be retrieved from this Backup servers.

- In our project, we are using three Backup Servers, and the encoded blocks of data are stored in this three Servers.

- For instance, if a file is divided into two blocks b1 and b2, then one block b1 in Server s1, block b2 in Server s2,etc

- Due to this, security of Backup servers is increased.

5. ADVANTAGES:

1. Efficient data Storage Allocation and Efficient Volume of Replication.

2. Higher reliability in which the data chunks are distributed on cloud servers.

3. Security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic secret sharing scheme in distributed storage systems.

4. No need of memorizing, managing, or sharing keys between the sender and receiver.

5. Hacking attacks like Masquerade attacks, Host based attacks and Network based attacks are avoided.

6. Large amount of data can be processed easily without hanging problem.

6. FUTURE SCOPE OF THIS PROJECT:

In future, specific algorithms will be developing to understand how certain applications store data i.e. linking it to Artificial Intelligence.

The traditional approach of convergent encryption cannot be suited in this secure deduplication as it is susceptible to brute-force attack. To overcome this drawback, modified version of convergent encryption can be used by introducing two approaches - domain separation and cryptographic tuning. This gives a better authorized deduplication approach.

Also, more techniques which would compress these encrypted block data into smaller size can be utilized for more efficient storage management.

7. CONCLUSIONS

Hence, as above mentioned we propose that this project will help efficient use of Storage provided by Cloud by Deduplicating data stored on it. Duplicate files or chunks of data will be stored as a single copy and this system will improve the reliability of data while achieving the confidentiality of the user's. The security of tag consistency and integrity were achieved.

ACKNOWLEDGEMENT

Here we express our sincere thanks to our project guide Ms. Pallavi Maladkar and HOD Prof M.S.Malkar for their constant support and for providing platform. We are also thankful to all other researchers for their publications.

REFERENCES

[1] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digi tal shadows, and biggest growth in the far east," http://www.emc.com/collateral/analyst reports/idcthe- digital-universe-in-2020.pdf, Dec 2012.

[2] M. O. Rabin, "Fingerprinting by random polynomials," Center for Research in Computing Technology, Harvard University, Tech. Rep. Tech. Report TR-CSE-03-01, 1981.

[3] 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 ICDCS, 2002, pp. 617–624.

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in USENIX Security Symposium, 2013.

[5] ——, "Message-locked encryption and secure deduplication," in EUROCRYPT, 2013, pp. 296–312.

[6] Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE - http://ieeexplore.ieee.org/document/7511769/

[7] Vikram V. Badge, Mrs. Rushali A. Deshmukh -Block Level Data Deduplication for Faster Cloud Backup - https://www.ijser.org/researchpaper/Block-Level-Data-Deduplication-for-Faster-Cloud-Backup.pdf

[8] Frederik Armknecht, Jens-Matthias Bohli, Ghassan O. Karame, Franck Youssef NEC Laboratories Europe 69115 Heidelberg, Germany - Transparent Data Deduplication in the Cloud.

[9] Golthi Tharunn, Gowtham Kommineni, Sarpella Sasank Varma, Akash Singh Verma of Electronics and Communication Department, GITAM University Rudraram, Hyderabad, Telangana, India - Data Deduplication in Cloud Storage.

[10] IBM - What is Cloud Computing: https://www.ibm.com/cloud/learn/what-is-cloud-computing.

[11] Technopedia - Cloud Computing : https://www.techopedia.com/definition/2/cloud-computing

[12] Investopedia - Cloud Computing : https://www.investopedia.com/terms/c/cloud-computing.asp

[13] Global Dots - Types of Cloud : http://www.globaldots.com/cloud-computing-types-of-cloud/

[14] How Data Deduplication Works by Enterprise Storage Guide - http://www.enterprisestorageguide.com/how-data-deduplication-works

[15] Naziya Tabassum, Prof. Roshani B. Talmale - Cryptographic Tuning to Minimize Storage Requirement on Cloud Using De-Duplication Mechanism - http://ijarcsse.com/Before_August_2017/docs/papers/Volume_6/5_May2016/V6I5-0331.pdf

[16] K Uma , E Jayabalan wrote - Improving Efficient Reliability Based Secure Deduplication Technique on Encrypted Big Data - https://www.ijircce.com/upload/2017/june/162_Improving.pdf