A Survey on Deduplication Checking In Cloud Computing

Prasad Chavan[1], Aishwarya Dupatre[2], Saiprasad Chalikwar[3] ,N. M. Waghdarikar[4] ,N. S. Kawathekar[5]

[1,2,3,4,5,] *Dept.of E&TC Engg., Smt.Kashibai Navale College of Engineering,Pune,Savitribai Phule Pune University, Pune*

[1]*prasadchavan631@gmail.com*

[2]*aishwarya.dupatre@gmail.com*

[3]*saiprasadchalikwar10@gmail.com*

[4]*nmwagdarikar@sinhgad.edu*

[5]*nachiket.kawathekar_skncoe@sinhgad.edu*

Abstract *- Cloud Computing has been one of the hottest buzzwords over the last few years but it is surprisingly known that the people have been using it for more than 10 years. Gmail, Facebook, Dropbox, Skype, PayPal, and Salesforce.com are all examples of cloud solutions which was not thinking about them in these terms. The main idea behind the cloud is that the information can be accessed over the internet without having any exhaustive familiarity of the communications used to enable it. The major services existing in Cloud computing is the Cloud storage. With the cloud storage, data can be stored on multiple third party servers which is not cared by the user and no one knows where exactly data saved. With the increase in size of the data every day, there is a need to handle, manage and mainly to store data, is a major problem faced by the people or organization. This article specifies about the study on space occupied by duplicate data over cloud. Where the data is increasing day by day , at the same time one thing to be noticeable that enough space at cloud is occupied by duplicate data so there is a need to check the data for de duplication at cloud before uploading.*

Keywords *- Data DE duplication, cloud, AES, MD5, Java, JSP & Servlet, etc*

I. INTRODUCTION

Cloud computing is the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing[1].Cloud computing provides various services in which data storage is the main cloud service. Cloud computing works behind the scene in our day to day activities such as to watch movies, play games, sending mails and listen to music etc, With Cloud computing, we can store, recover and backup data, create new applications, deliver software on demand,

host websites and so on. Whenever there is a demand, user can access the services of cloud dynamically via internet[2]. There are three types of cloud computing services models namely SaaS (Software as a service), PaaS (Platform as a service) and IaaS (Infrastructure as a service). SaaS is a cloud computing offering that provides users with access to a vendor's cloud-based software. PaaS is a cloud computing offering that provides users a cloud environment in which they can develop, manage and deliver applications. IaaS is a cloud computing offering in which a vendor provides users access to computing resources such as servers, storage, and networking[3]. Big data storage is a compute-and-storage architecture that collects and manages large data sets and enables real-time data analytics. As the technology is mounting, the size of data is also growing accordingly. So people are living in the world of big data. The term big data refers to the dataset of huge size which are incapable to store in typical database[4]. Big data often lacks structure and comes from various sources, making it a poor fit for processing with a relational database. Cloud Data Storage is made out of thousands of distributed storage gadgets grouped by system, disseminated document frameworks and other stockpiling middleware to give distributed storage administration to clients. The normal structure of CDS incorporates capacity asset pool, circulated record framework, benefit level assentions (SLAs), and administration interfaces, and so on. Comprehensively, they can be isolated by physical and sensible capacities limits and connections to give more compatibilities and communications. Compact discs is having a tendency to joined with CDSS, which will give progressively vigorous security. Cloud storage is one of the primary use of cloud computing. With the cloud storage, data is stored on multiple third party servers, rather than on the dedicated servers used in traditional networked data storage. When storing data, the user sees a virtual server which is called that it appears as if the data is stored. But it does not exist in reality which is just a pseudonym used to reference virtual space carved out of the cloud. In reality, the user's data could be stored on any one or more of computers used to create the cloud [5]. The basic level in cloud storage system is that it needs one data server connected to the internet. A client sends copies of files over internet to the data server, which then records the information. When client wishes to retrieve the information, he or she accesses the data server through a web based interface. The server then either sends the files back to the client or allows the client to access and manipulate the files on the server itself.

II. RELATED WORK

Data in most companies are processed by Hadoop by submitting the jobs to Master. The Master distributes the job to its cluster and process map and reduces tasks sequentially. But nowadays the growing data need and the competition between Service Providers leads to the increased submission of jobs to the Master. This Concurrent job submission on Hadoop forces us to do Scheduling on Hadoop Cluster so that the response time will be accept- able for each job. In this Deduplication techniques are most widely employed to backup data and minimize network and storage overhead by detecting and eliminating redundancy among data. So which is crucial for eliminating dupli- cate copies of identical data in order to save storage space and network band- width? We present an attribute-based storage system with secure deduplication in a hybrid cloud setting, using public cloud and private cloud. Where a pri- vate cloud is responsible for duplicate detection and a public cloud manages the storage. Instead of keeping multiple data copies with the same content, the system eliminates redundant data by keeping only one physical copy and re- ferring other redundant data to that copy. Each such copy can be defined based on user access policies. In this user will upload the file with access policies and then file type question with answer. Then same file with different access policies to set the particular file to replace the reference. Where a users private key is associated with an attribute set, a message is encrypted under an access policy over a set of attributes, and a user can decrypt a ciphertext with his/her private key if his/her set of attributes satisfies the access policy associated with this ciphertex.

2. "S. Kalaivani, Secure Data Sharing in Cloud Computing using Revocable Storage Identity Based Encryption. 2017

Nowadays regularly use cloud services in our daily life. There are various services provided by cloud such as a service, Platform as a service, and Infras- tructure asa service. The used to keep our data,documents, and files on cloud. The data that store may be Personal, Private, secret data. So must be very sure that whatever the cloud service we use that must be secure. Cloud computing Provides number of services to client over internet. Storage service isone ofthe important services that people used now days for storing data on network so that they can access their data from anywhere and anytime. With the benefit of storage service there is an issue of security. To overcome security problem the proposed system contain two levels of security and to reduce the unwanted storage space de-duplication[1,2] technique is involved. To increase the level of security one technique is a session password.Session passwords can be used only once and every time a new password is generated.To protect the confiden- tiality of sensitive data while supporting de-duplication[1,2]the convergent en- cryption technique has been proposed to encrypt the data before outsourcing, Symmetrickey algorithm uses same key for both encryption and decryption. In this paper, I will focus on session based authentication for both encryptions for files and duplication check for reduce space of storage on cloud.

III. PROPOSED SYSTEM

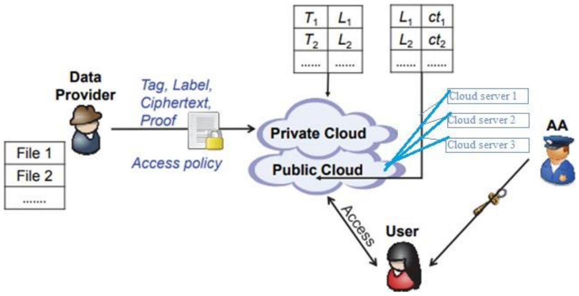


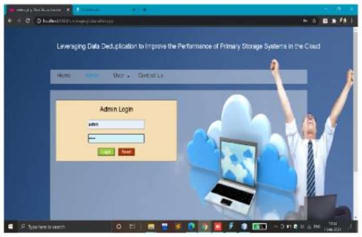
Fig: Proposed System

An attribute-based storage system supporting secure de-duplication. Our storage system is built under a hybrid cloud architecture, where a private cloud manipulates the computation and a public cloud manages the storage. Attribute based storage system supporting secure de duplication of encrypted data in the cloud, in which the cloud will not store a file more than once even though it may receive multiple copies of the same file encrypted under different access policies. The Attribute Authority issues every user a decryption key associated with the set of attributes. The attribute based storage system check the duplication of the file. The duplication is not occur, the file is stored. If the duplication is occurring, the attribute authority changes the ownership permission. In this system utilizing client accreditations to check the confirmation of the client. In that cases cloud is available two sort of cloud such private cloud and open cloud. In private cloud store the client accreditation and in the open cloud client information present out. The system have utilized a half and half cloud construction modeling as a part of proposed. In this system have to need to mind the file name in record information duplication and information DE duplication is checked at the square level. On the other hand, client needs to recover his information or download the information record he have to download both of the document from the cloud server this will prompts perform the operation on the same record this abuses the security of the distributed storage. Division and Replication of Data in the Cloud for Optimal Performance and Security (DROPS) that collectively approaches the security and performance issues. In this project, DROPS methodology, divide a file into fragments, and replicate the fragmented data over the cloud nodes. Each of the nodes stores only a single fragment of a particular data file that ensures that even in case of a successful attack, no meaning-ful information is revealed to the attacker.

Algorithm:
MD5
Step 1.Append Padding Bits. The message is ”padded” (extended) so that its length (in bits) is congruent to 448, modulo 512. ...
Step 2. Append Length. ...
Step 3. Initialize MD Buffer. ...
Step 4. Process Message in 16-Word Blocks. ...
Step 5. Output

In cryptography, MD5 (Message-Digest algorithm 5) is a widely used cryptographic hash function with a 128-bit hash value. As an Internet standard (RFC 1321), MD5 has been employed in a wide variety of security applications, and is also commonly used to check the integrity of files. An MD5 hash is typically expressed as a 32 digit hexadecimal number.

Experimental Results:



Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud

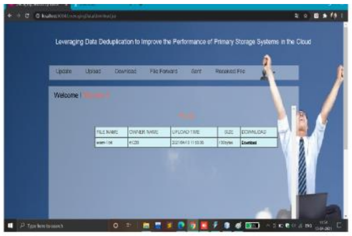
Admin Login

admin

Login Page

Welcome ! Admin

Home Page



Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud

Welcome ! [illegible]

FILE NAME	OWNER NAME	UPLOAD TIME	SIZE	DOWNLOAD
[illegible]	[illegible]	[illegible]	130bytes	Download

ISSN: 2233-7853 IJFGCN

File Upload /Download Details

IV. CONCLUSION:

Thus we are going to develop a system for secure deduplication in cloud computing. Here the files will be first checked either they have been already uploaded or not , and if any file is already uploaded then it will not be uploaded again. This system will help to improve the efficiency of the cloud storage system. It will solve the problem of availability of storage space to great extent.

REFERENCES

[1] https://aws.amazon.com/what-is-cloud-computing/
[2] Dr.P.Sujatha and Dr.P.SriPriya, "Security Threats and Preventive Mechanisms in Cloud Computing ", JASC: Journal of Applied Science and Computations Volume V, Issue XII, December/2018 ISSN NO: 1076-5131.
[3] K.Sharmila S. Borgia Anne Catherine Sreeja V.S, "A comprehensive Study of Data Masking Techniques on cloud", International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018, 3719-3727.
[4] K. Sharmila and Dr.S.A.Vethamanickam, " MRK-SVM: An Effective Technique for Big Data In Health Care Sector", International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016,ISSN 2229-5518.
[5] Wassim Itani Ayman Kayssi Ali Chehab, “Privacy as a Service: Privacy-Aware Data Storage and Processing in Cloud Computing Architectures”, Eighth IEEE International Conference on Dependable, 2009.
[6] Takahiro Hirofuchi, Hidemoto Nakada, Hirotaka Ogawa, Satoshi Itoh,
Satoshi Sekiguchi. 2009. A live storage migration mechanism over wan and its performance evaluation. Proceedings of the 3rd international workshop on Virtualization technologies indistributed computing, Barcelona, Spain, 2009, 67-74.
[7] FalconStor Software, Inc. 2009. Demystifying Data Reduplic ation: Choosing the Best Solution. http://www.ipexpo.co.uk/content/download/20646/353747/file/DemystifyingDataDedupe_WP.pd f, White Paper, 2009-10-14, 1-4.
[8] Steve Lesem. 2009. Cloud Storage and The Innovator's Dilemma. http://cloudstoragestrategy.com/cloud-ecosystem/, July 19, 2009.
[9] Storage Networking Industry Association.Cloud Storage Reference Model,Jun.2009.
[10] R. Arokia Paul Rajan, S. Shanmugapriyaa “Evolution of Cloud Storage as Cloud Computing Infrastructure Service” IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 1, Issue 1 (May-June 2012), PP 38-45.