International Journal of Computer Science and Engineering Communications
Vol.3, Issue 3, 2015, Page.1086-1091
ISSN: 2347–8586
www.scientistlink.org

Reliable Secure Data Storage in the Cloud Environments and De duplication

MD. Rafeeq[1], C. Sunil Kumar[2]
[1]Associate Professor, Department of CSE,CMRTC, Hyderabad, India.
[2] Professor, SNIST, Hyderabad, India.

ABSTRACT
Recent Developments in the organizations have witnessed the trend of leveraging cloud-based services for large scale content storage, processing, and distribution. Security and privacy are among Top concerns for the public cloud environments. Towards these security challenges a reliable propose and implementation on Open Stack Swift, a new client-side de duplication scheme for securely storing and sharing outsourced data via the public cloud. Security issues, requirements and challenges that cloud service providers (CSP) face during cloud engineering. Mainly this proposal has twofold. First, it ensures better confidentiality towards unauthorized users. Second, by integrating access rights in meta data file. That is, every client computes as per data key to encrypt the data that he intends to store in the cloud.

Key words- Content Storage, processing and Distribution, Security and privacy, De Duplication.

I. INTRODUCTION

Data de duplication refers to the elimination of redundant data. De duplication algorithms identify and delete duplicate, leaving only one copy (or 'single instance') of the data to be stored. However, indexing of all data is still retained should that data ever be needed. Towards this security, a new client-side de duplication scheme for securely storing and sharing outsourced data via the public cloud, an authorized user can decipher an encrypted file only with his private key. Cloud service providers (CSP) (e.g. Microsoft, Google, Amazon, Salesforce.com, Go Grid) re leveraging virtualization technologies combined with self-service capabilities for computing resources via the Internet. Today, enterprises are looking toward cloud computing horizons to expand their on-premises infrastructure, but most cannot afford the risk of compromising the security of their applications and data. Assume that there is an established secure channel between the client and the CSP. This secure channel supports mutual authentication and data confidentiality and integrity. Hence, after successfully. Authenticating with the CSP, these Cloud users share the same resources in a multi-tenant environment.

International Data Corporation (IDC) conducted a survey [1] (see Fig.1.) of 263 IT executives and their line-of-business colleagues to gauge their opinions and understand their companies' use of IT cloud services. Security ranked first as the greatest challenge or issue of cloud computing.

Corporations and individuals are concerned about how Security and compliance integrity can be maintained in this new environment. Even more concerning, though, is the corporations that are jumping to cloud computing while being oblivious to the implications of putting critical applications and data in the cloud. Moving critical applications and sensitive data to a public and shared cloud environment is a major concern for corporations that are moving beyond their data center's network perimeter defense. To alleviate these concerns, a cloud solution provider must ensure that customers can continue to have the same security and privacy controls over their applications and services, provide evidence to these customers that their organization and customers are secure and they can meet their service-level agreements, and show how can they prove compliance to their auditors.

Reliability

Servers in the cloud have the same problems as your own resident servers. The cloud servers also experience downtimes and slowdowns, what the difference is that users have a higher dependent on cloud service provider (CSP) in the model of cloud computing. There is a big difference in the CSP's service model, once you select a particular CSP, you may be locked-in, thus bring a potential business secure risk.

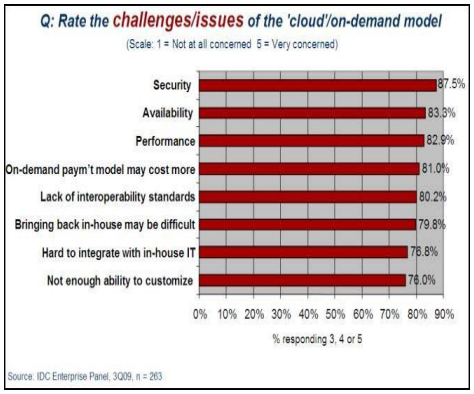


Fig. 1. Results of IDC ranking security challenges

Security: Where is your data more secure, on your local hard driver or on high security servers in the cloud? Some argue that customer data is more secure when managed internally, while others argue that cloud providers have a strong incentive to maintain trust and as such employ a higher level of security. Virtually any server, and there are the statistics that show that one-third of breaches result from stolen or lost laptops and other devices and from employees' accidentally exposing data on the Internet, with nearly 16 percent due to insider theft [8]. This professional paper discusses security and reliability issues, challenges, and recommends control objectives to technical and business community. It also highly recommends OVF standard as vendor and platform independent, open, secure, portable, efficient and extensible format for the packaging and distribution of software to be run in virtual machines.

II. RELIABILITY IN THE CLOUD

To advance cloud computing, the community must take proactive measures to ensure security. The Berkeley paper's solution is the data encryption. Before storing it at virtual location, encrypt the data with your own keys and make sure that a vendor is ready for security certifications and external audits. Identity management, access control, reporting of security incidents, personnel and physical layer management should be evaluated before you select a CSP. And you should minimize personal information sent to and stored in the cloud. CSP should maximize the user control and provide feedback. Organizations need to run applications and data transfer in their own private cloud and then transmute it into public cloud. While there are many legal issues exist in the cloud computing.

Data Life Cycle

Data life cycle refers to the entire process from generation to destruction of the data. The data life cycle is divided into seven stages. See the figure below:

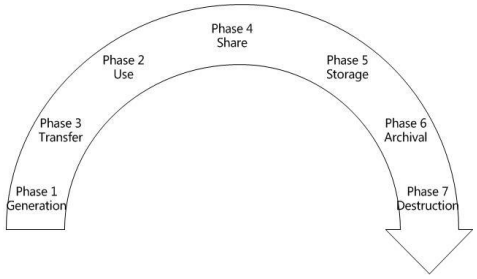


Figure 2. Data life cycle

1. Data Generation

Data generation is involved in the data ownership. In the traditional IT environment, usually users or organizations own and manage the data. But if data is to be migrated into cloud, it should be considered that how to maintain the data ownership.

2. Transfer

Within the enterprise boundaries, data transmission usually does not require encryption, or just have a simple data encryption measure. For data transmission across enterprise boundaries, both data confidentiality and integrity should be ensured in order to prevent data from being tapped and tampered with by unauthorized users.

3. Use

For the static data using a simple storage service, such as Amazon S3, data encryption is feasible. However, for the static data used by cloud-based applications in PaaS or SaaS model, data encryption in many cases is not feasible.

4. Share

Data sharing is expanding the use range of the data and renders data permissions more complex. The data owners can authorize the data access to one party, and in turn the party can further share the data to another party without the consent of the data owners. Therefore, during data sharing, especially when shared with a third party, the data owners need to consider whether the third party continues to maintain the original protection measures and usage restrictions.

5. Storage

The data in the cloud may be divided into: (1) The data in IaaS environment, such as Amazon's Simple Storage Service; (2) The data in PaaS or SaaS environment related to cloudbased applications. The data stored in the cloud storages is similar with the ones stored in other places and needs to consider three aspects of information security: confidentiality, integrity and availability.

6. Archival

Archiving for data focuses on the storage media, whether to provide off-site storage and storage duration. If the data is stored on portable media and then the media is out of control, the data are likely to take the risk of leakage. needs to consider the use of both encryption algorithm and key strength.

7. Destruction

When the data is no longer required, whether it has been completely destroyed? Due to the physical characteristics of storage medium, the data deleted may still exist and can be restored.

CURRENT SECURITY SOLUTIONS FOR DATA:

IBM developed a ***fully homomorphism*** encryption scheme in June 2009. This scheme allows data to be processed without being decrypted [2]. Roy I and Ramadan HE applied decentralized information flow control (DIFC) and differential privacy protection technology into data generation and calculation stages in cloud and put forth a privacy protection technology into data generation and calculation stages in cloud and put forth a privacy protection system called airavat [3]. This system can prevent privacy leakage without authorization in Map-Reduce computing process. A key problem for data encryption solutions is key management. On the one hand, the users have not enough expertise to manage their keys. On the other hand, the cloud service providers need to maintain a large number of user keys. The Organization for the Advancement of Structured Information Standards (OASIS) Key Management Interoperability Protocol (KMIP) is trying to solve such issues[4]. In the data storage and use stages, Mowbray proposed a client-based privacy management tool [7]. It provides a user centric trust model to help users to control the storage and use of their sensitive information in the cloud. Munts-Mulero discussed the problems that existing privacy protection technologies (such as K anonymous, Graph Anonymization, and data pre-processing methods) faced when applied to large data and analyzed current solutions [8].

The challenge of data privacy is sharing data while protecting personal privacy information. Randike Gajanayake proposed a privacy protection framework based on information accountability (IA)components [5]. The IA agent can identify the users who are accessing information and the types of information they use. When inappropriate misuse is detected, the agent defines a set of methods to hold the users accountable for misuse. methods to hold the users accountable for misuse. About data destruction, U.S. Department of Defense (DoD)5220.22-M (the National Industrial Security Program Operating Manual) shows two approved methods of data (destruction) security, but it does not provide any specific requirements for how these two methods are to be achieved.

Application Related Security Issues

Application security refers to using system resources such as the software and hardware to ensure security of applications, which guards against intrusion from the malicious attackers. At present, one general type of attack is disguising as a trusted user and then get the full access of the system which is tricked. The main reason is that the network level security policies are obsolete, which allow only authorized user can access assigned IP address. In cloud computing, recent technologies make it quite possible to impersonate an authorized user and breach the data stealthily.

Cloud browser security:

In a SaaS model, the client's computation tasks are outsourced to the remote servers. The client system is used only for IO, receiving and sending commands to the cloud. The web browser is an universal client application which satisfies this demand. In this context, the browser security is especially important in cloud computing[6].

VM security and threats:

It 's difficult to efficiently manage many VMs running on a same physical machine. Each customer is responsible for his own VM, updating and patching the operating system and all the software on his own, which leads to security holes that can be exploited by an attacker. It is not impossible for an hacker to attack a guest OS, then manage to run a piece of malicious code on encryption.

Distributed denial of service (DDoS) attacks in cloud:

Generally, DDoS may be more dangerous than denial of service (DoS) in terms of denying important services on a server by flooding the target server with a large number of packets which can not be processed by it. Unlike Dos, DDos attack is relayed from different dynamic networks which have already been compromised. In the context of the cloud computing, the situation is similar but even more worse because the attacker could use more VMs to launch attack in the environment of an IaaS model[4].

III. Data De duplication Methods

De duplication is able to reduce the required bandwidth and storage capacity, since only the unique data is stored. For example, a typical email system might contain 100 instances of the same 1 MB fi le attachment. If the email platform is backed up or archived, all 100 instances are saved, requiring 100 MB storage space. With data de duplication, only one instance of the attachment is actually stored; each subsequent instance is just referenced back to the one saved copy. In this example, a 100 MB storage and bandwidth demand could be reduced to only 1 MB.

Server Side versus Client Side

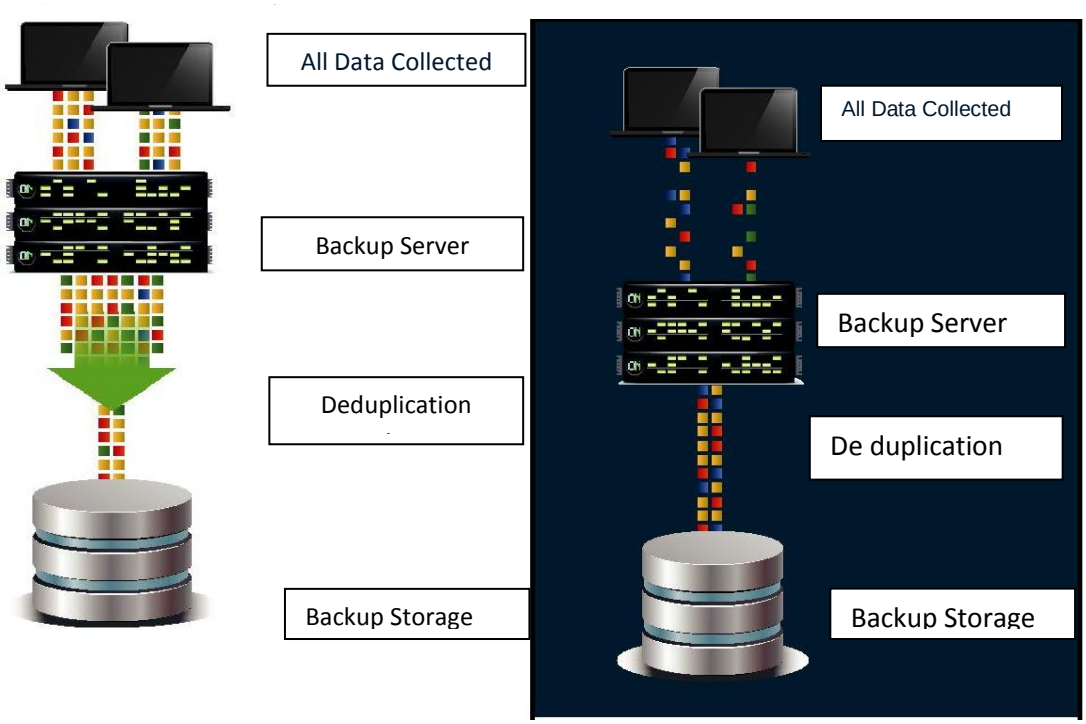
In order to prevent data breach on lost or stolen devices, in Sync provides remote wipe Server side based de duplication method acts on the data on the server. In this case, the client is una- ected and does not benefi t from any de duplication. The de duplication engine can be embedded in the hardware array, which can be used as NAS/SAN device with de duplication capabilities. Alternatively it can also be offered as an independent software or hardware appliance which acts as intermediary between backup server and storage arrays. In both cases, this method does not decrease the amount of data transmitted and provides no improvements on *bandwidth utilization it improves only the storage utilization.* The client-side de duplication method acts on the data at the client i.e. before it is moved to the server. A de duplication-aware backup agent is installed on the client, the agent backs up only unique data. This approach results in improved *bandwidth and storage utilization*. However, this method imposes additional computational load on the backup client.

File versus Sub-file Level Data De duplication

The duplicate removal algorithm can be applied on full file or sub-file levels. File-level duplicates can be easily eliminated by calculating a single checksum of the complete file data and comparing it against existing checksums of backed-up files.

This approach is simple and fast, but the extent of de duplication is very small, as it does not address the problem of duplicate content found inside different files or data-sets (e.g. emails).The sub-file level de duplication technique breaks the file into smaller fixed or variable size blocks, and then uses standard hash-based algorithms to find similar blocks.

International Journal of Computer Science and Engineering Communications
Vol.3, Issue 3, 2015, Page.1086-1091
ISSN: 2347–8586
www.scientistlink.org



Block-based De duplication

The block-based de duplication algorithms work the following way: the de duplication engine looks at a sequence of data, segment sit into variable length blocks, and seeks blocks that are repeated. The engine stores a pointer to the original block instead of storing the duplicate block again. There are two main types of the block-based approach.

Fixed-length block approach, as the name suggests, divides the files into fixed-length blocks and uses simple checksum (MD5/SHA etc.) based approach to fi nd duplicates.

Although it's possible to look for repeated blocks, the approach provides very limited effectiveness since the primary opportunity for data reduction is in finding duplicate blocks in two transmitted datasets that are made up mostly - but not completely - of the same data segments. For example, similar data blocks may be present at different o- sets in two different datasets. In other words the block boundary of similar data may be different. This is very common when some bytes are inserted in a fi le, and when the changed file processes again and divides into fixed-length blocks, all blocks appear to have changed. Therefore, two datasets with a small amount of difference are likely to have very few identical fixed-length blocks. Variable-Length Data Segment technology divides the data stream into variable length data segments using a methodology that can find the same block boundaries in different locations and contexts. This allows the boundaries to "fl oat" within the data stream so that changes in one part of the dataset have little or no impact on the boundaries in other locations of the dataset. Through this method, duplicate data segments can be found at different locations inside a file or between different files created by same/different application.

Limitations of Block-Based De duplication

1. The block size (fixed or floating) used to determine data boundary is usually a “best” guess, & hence may not completely coincide with application’s actual block size.

2. Different applications have different ways of writing on-disk data, block based algorithm will often fail to detect identical blocks across different application file types (e.g. the same block of text stored in MS Word file and in a .PST email file).

3. Applications like Microsoft Outlook and Once use a complex database based on disk data structure which “stamp” each block with aunique header and footer, further complicating the task of finding duplicate blocks of data.

IV. CONCLUSION

Cloud computing faces many of the same challenges as other information and network technologies: performance, security, resiliency, interoperability, data migration, and transition from legacy systems. We find that whole-file de duplication together with sparseness is a highly efficient means of lowering storage consumption, even in a backup scenario. It approaches the effectiveness of conventional de duplication at a much lower cost in performance and complexity. The environment we studied, despite being homogeneous, shows a large diversity in file system and file sizes. These challenges, the increase in unstructured files, and an ever-deepening and more populated name space pose significant challenge for future file system designs. However, at least one problem – that of file fragmentation, appears to be solved, provided that a machine has periods of in activity which de fragmentation can be run.

-Security issues indicate potential problems which might arise.
-Security standards offer some kind of security templates which cloud service providers (CSP) could obey.
-The most promising standard for the future would be OVF format which promises creation of new business models that will allow companies to sell a single product on premises, on demand, or in a hybrid deployment model.

ACKNOWLEDGEMENTS

I would like to extend thanks to the Director and management of CMR Technical Campus , who so generously supported to publish this paper and I thank Dr. C. Sunil Kumar Professor, for his valuable suggestions and supervision.

REFERENCES

1. http://en.wikipedia.org/wiki/Cloud_computing.
2. Rich Maggiani, solari communication. “Cloud computing is changing how we communicate”.
3..Randolph Barr, Qualys Inc, “How to gain comfort in losing control to the cloud”.
4. Tharam Dillon, Chen Wu,Elizabeth Chang,2010 24th IEEE International Conference on Advanced Information Networking and Applications ,“ Cloud computing: issues andchallenges”.
5. International Data Corporation, 2/idc
6. CloudS ecurity Alliance. http://www.cloudsecurityalliance.org.
7. Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing,V2.1,
8. A survey on security issues in service delivery models of cloud computing.Journal of Network and Computer Applications 34(2011)1-11.