ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Literature Survey on Deduplication of Health Care Data in Cloud Environment.

R Shiny Sharon*, and R Joseph Manoj.

Department of IT, St. Joseph's College of Engineering, Chennai, Tamil Nadu, India.

ABSTRACT

Cloud computing is an Internet-based computing which provides shared processing resources and data to computers and devices on demand. Cloud computing provide users with various capabilities to process their data in third party data centres. It has become a highly demanded service due to computing power, cheap cost for services offered, high performance and availability. Three benefits of cloud include self-servicing provisioning, elasticity, pay per use. Cloud computing services can be private, public and hybrid. Deduplication is a technique used to store files in the cloud. E- Healthcare system plays a major role in the society. It monitors the health condition and helps in giving appropriate medical treatments. This system aims at gathering and storing patient's details and sharing health related information. Deduplication is used in E-Healthcare system to avoid duplicated copies of patient data. Deduplication is a method used for eliminating redundant data and allows only one copy of the file to be stored. The more efficient use of disk space also allows for longer disk retention periods. Deduplication does not allow more than one copy of the file to be saved as it occupies most of the memory space. Deduplication initially compares the title of the available files and the file that is to be saved. If the title does not match with the available data, it then compares with the contents of the files that are available in cloud, if the content is also unique then the file is allowed to be stored in cloud. If either the title or the content is same then the file will be blocked from saving.

Keywords: Cloud Computing, E-Healthcare, Deduplication,

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Cloud computing is one of the emerging technologies in the world of computers. It is a set of hardware and software that collectively work to deliver services on a wide range over a particular network. Users can access files and other data applications from any data that uses the internet connection. Cloud computing aims at providing many aspects of computing to the end users through online servicing. The cloud computing are used in entertainment, medical, military operations, security issues, business and finance etc.

There are different types of clouds that you can subscribe to depending on your needs. As a home user or small business owner, you will most likely use public cloud services. 1. Public Cloud - A public cloud can be accessed by any subscriber with an internet connection and access to the cloud space. 2. Private Cloud - A private cloud is established for a specific group or organization and limits access to just that group. 3. Community Cloud - A community cloud is shared among two or more organizations that have similar cloud requirements. 4. Hybrid Cloud - A hybrid cloud is essentially a combination of at least two clouds, where the clouds included are a mixture of public, private, or community.

APPLICATIONS OF CLOUD COMPUTING

Cloud in field of Business

The business delivery model provides a user experience by which hardware, software and network resources are optimally leveraged to provide innovative services over the Web, and servers are provisioned in accordance with the logical needs of the service using advanced, automated tools. Companies that employ traditional data centre management practices know that making IT resources available to an end user can be time intensive. It involves many steps, such as procuring hardware; finding raised floor space and sufficient power and cooling; allocating administrators to install operating systems, middleware and software; provisioning the network; and Securing the environment Cloud computing reduces the need for physical infrastructure and scales to provide efficient use of energy based on real-time requirements. Cloud computing provides smaller companies with the ability to use enterprise applications that they couldn't otherwise afford. File sharing and workflow in the cloud increases collaboration and communications efficiency. Cloud computing also provides security on a higher scale and also provides cost predictability which allows the companies to budget them at their initial stages.

Cloud in field of Entertainment

Most people on the internet are longing for entertainment. Cloud computing is the perfect place for reaching to a varied consumer base. Cloud-based entertainment can reach any device be it TV, mobile or any other form. Cloud Computing provides on demand storage and compute power to be billed in a pay-per-use basis, comes as a perfect strategic fit to solve the puzzle of ODE (On Demand Entertainment). Cloud Computing can provide a solution to the issue of huge requirements in computing and storage. Cloud providers are aware that security and content protection are the key issues for media world. Cloud computing over Entertainment provides a new business insights into how cloud computing can extend their core enterprise capabilities and it will provide further future guidance to attain the goals. Moreover media companies cannot simply raise their prices or sell more advertising to attain profit goals and there is not much left from cutting operating expenses. Everything purely relies on the effectiveness of digital supply chains. For these reasons media companies are looking into cloud for enhancing their digital supply chains. Cloud also enables the ability to distribute more B2B and B2C contents to more consumers for the least investments.

Cloud in field of Banking

Cloud computing eliminates the need for having a separate banking portal and client database for every location. Despite being the slow adapters in the industry, banking as concerns with security and reliability, financial institutions are quickly resorting to cloud-based services to achieve lowered total cost of ownership (TCO). Over the years financial institutions typically have been consumers of cloud-based solutions across generic and non-core services like virtualization, data centre consolidation, and storage and disaster recovery. Many financial institutions are either planning or have implemented in-house private clouds for sensitive consumer data and are utilizing the public cloud for generic services. As cloud computing capabilities

mature and become more reliable, multi-tenancy and hybrid cloud models will drive increased adoption of cloud-based solutions that are focused on core services and achieve cost efficiencies and scalability.

CHALLENGES IN CLOUD COMPUTING AND DUPLICATE DATA

Security issues

The valuable enterprise data will reside outside the firewall that will raise serious problems. Attacks made towards cloud will affect multiple clients even if only one site is attacked. These risks can be mitigated using security applications, encrypted file systems, data loss software and other security hardware to track down the unusual behaviours across servers.

Interoperability and portability issues

Businesses should have the leverage of migrating in and out of the cloud and switching providers whenever they want, and there should be no lock-in period. Cloud computing services should have the capability to integrate smoothly with the premises of IT.

Reliability and Availability

Cloud providers still lack round-the-clock service, this results in frequent outages. It is important to monitor the service being provided using internal or third-party tools. It is vital to have plans to supervise usage, SLAs, performance, robustness, and business dependency of these services.

Duplicate Copies of Data

Duplicate copies of data will lead to wastage of memory space and chaos. Redundancy of data will be minimized and the throughput will be reduced. This will become a major problem while retrieving the particular data from a large pool of data set.

RELATED STUDY

Private data deduplication technique for storing private and personalized data was introduced and formalized by Wee Keong Ng SCE, Yonggang Wen SCE, and Huafei Zhu [1]. Where a private data deduplication protocol allows a client, who holds a private data which proves to a server who holds a summary string of the data that he/she is the owner of that data without revealing further information to the server. The security of the private data is formalized using private data deduplication techniques. This protocol is the first data deduplication for private data.

An architecture in DupLESS: Server-Aided Encryption for Deduplicated Storage system was described by Mihir Bellare and Sriram Keelveedhi [2] which provides secured deduplicated storage resisting brute-force attacks, and realize it in a system called DupLESS (Duplicate less Encryption for Simple Storage). It provides more secure, easily-deployed solution for encryption that supports deduplication. In DupLESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. Clients authenticate themselves to the Key-Server, but do not leak any information about their data to it. As long as the Key-Server remains in- accessible to attackers, we ensure high security. It enables clients to store the encrypted data and achieves strong confidentiality of data. DupLESS is more of feature compatible with the API commands in the system. The Deduplication subroutine enables fine grained control over the files and will be deduplicated, for exception the data in the personal files will not be deduplicated. Thus DupLESS provides security that is significantly better than current, convergent encryption based deduplicated encryption architectures.

Deduplication is known to effectively eliminate duplicates for Virtual Machine (VM) image storage which was explained in RevDedup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backups by Chun-Ho Ng and Patrick P. C. Lee in June 28, 2013 [3]. Where it introduces fragmentation that will degrade read performance. RevDedup is introduced here, which is a deduplication system that optimizes reads to latest VM image backups is using an idea called reverse deduplication. RevDedup removes duplicates from old data, RevDedup achieves high deduplication efficiency with around 97% of saving, and high backup and

read throughput on the order of 1GB/s. Many backup solutions are made by disk-spaced storage systems which has better I/O performance than other traditional storage systems. Deduplication is mainly studied in content-addressable backup systems. It is also shown to provide space-efficient VM image storage given that VM images have significant content similarities. Here Deduplication mainly focuses on optimizing storage efficiency and performance. RevDedup exploits content similarities of VM images using a hybrid of inline and out-of-order deduplication approaches. It applies coarse-grained global deduplication (inline) to different VMs and removes any duplicates on the path, and further applies fine-grained reverse deduplication (out-of-order) to different backup versions of the same VM and removes any duplicates from old backup versions. Threshold-based block removal mechanism is used, that combines hole-punching to remove duplicate blocks of old backup versions and segment compaction to compact data segments without duplicate blocks to reclaim contiguous space.

A novel encryption scheme that guarantees semantic security for unpopular data and provides weaker security and better storage and bandwidth benefits for popular data was introduced in A Secure Data Deduplication Scheme for Cloud Storage by Jan Stanek Alessandro Sorniotti, Elli Andreoulaki, and Lukas Kencl [4]. Data deduplication is made effective for popular data, whereas semantic security encryption protects the unpopular data. The effectiveness of storage efficiency functions such as compressions and deduplication is and objective for both storage provider and customer. High compression levels and deduplication ratios allow optimal usage of the resources of the storage provider and also lowers the cost for its users. Data deduplication proves that multiple uploads for the same content only consumes the network bandwidth and storage space for single upload. Deduplication is continuously used by many cloud providers as well as various cloud services to eliminate duplicated data. With the growing data size of cloud computing, a reduction in data volumes could help providers reducing the costs of running large storage system and saving energy consumption.

So data deduplication techniques have been brought to improve storage efficiency in cloud storages which was explained in Dynamic Data Deduplication in Cloud Storage by Waraporn Leesakul, Paul Townend [5].With the dynamic nature of data in cloud storage, data usage in cloud changes overtime, some data chunks may be read frequently in period of time, but may not be used in another time period. A dynamic deduplication scheme for cloud storage, which aims to improve storage efficiency and maintaining redundancy for fault tolerance is used. Data deduplication is a technique whose objective is to improve storage efficiency. With the aim to reduce storage space, in traditional deduplication systems, duplicated data chunks identify and store only one replica of the data in storage. Logical pointers are created for other copies instead of storing redundant data. Deduplication can reduce both storage space and network bandwidth. Current data deduplication mechanisms in cloud storage are static schemes applied agnostically to all data scenarios. Deduplication in cloud storages requires a dynamic scheme which has the ability to adapt to various access patterns and changes the user behaviour in cloud storages.

Cloud storage services commonly use deduplication technique for eliminating redundant data by storing only a single copy of data of each file of data block which was explained in Side channels in cloud services, the case of deduplication in cloud storage by Danny Harnik [6]. Privacy implication of cross-user deduplication is used in the system and it will be demonstrated as a side channel which reveals the information about the contents of the files of other users. High savings are offered by cross-user deduplication. Data deduplication strategies are categorized into two types they are 1) File-level deduplication 2) Block-level deduplication. In file-level deduplication only single copy of the file is stored and two or more files are identified as identical if they have the same hash value. In Block-level deduplication the data file is segmented into blocks and only single block will be stored. The system could use either fixed size block or variable sized blocks. The effectiveness of deduplication depends on multiple factors such the type of data, the retention period and the number of users. By applying this source based deduplication approach, client will be able to easily identify whether a certain file or block is deduplicated. This can be done by either examining the amount of data transferred over the network, or by observing the log of the storage software.

CONCLUSION

Privacy preserving techniques for securing the PHI such as Layered model of access structure which will solve the problem of multiple hierarchical files sharing. FH-CP-ABE is to be implemented which has low

storage cost and computation complexity in terms of encryption and decryption. Deduplication allows only a single instance of a file to be save which saves memory wastage and time will be implemented.

REFERENCES

[1] "Private Data Deduplication Protocols in Cloud Storage" Wee Keong Ng SCE, Yonggang Wen SCE, Huafei Zhu.

[2] "DupLESS: Server-Aided Encryption for Deduplicated Storage" Mihir Bellare and Sriram Keelveedhi, University of California, San Diego; Thomas Ristenpart, University of Wisconsin—Madison.

[3] "RevDedup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backups" Chun-Ho Ng and Patrick P. C. Lee The Chinese University of Hong Kong, Hong Kong Technical Report June 28, 2013.

[4] "A secure data deduplication scheme for cloud storage" Jan Stanek Alessandro Sorniotti, Elli Andreoulaki, Lukas Kencl.

[5] "Dynamic Data Deduplication in Cloud Storage" Waraporn Leesakul, Paul Townend, Jie Xu School of Computing University of Leeds, Leeds, LS2 9JT United Kingdom.

[6] "Side channels in cloud services, the case of deduplication in cloud storage" Danny Harnik IBM Haifa Research Lab Benny Pinkas Bar Ilan University Alexandra Shulman-Peleg IBM Haifa Research Lab.

[7] "Memory Deduplication as a Threat to the Guest OS" Kuniyasu Suzaki, Kengo Iijima, Toshiki Yagi, Cyrille Artho National Institute of Advanced Industrial Science and Technology.

[8] "Hierarchical Attribute-Based Encryption for Fine-Grained Access Control in Cloud Storage Services"Guojun Wang, Qin Liu School of Information Science and Engineering Central South University Changsha, Hunan Province, P. R. China, 410083 Jie Wu Dept. of Computer and Information Sciences Temple University Philadelphia, PA 19122, USA.

[9] "Scalable and Secure Sharing of Personal Health Records in Cloud Computing using Attribute-based Encryption" Ming Li Member, IEEE, Shucheng Yu, Member, IEEE, Yao Zheng, Student Member, IEEE, Kui Ren, Senior Member, IEEE, and Wenjing Lou, Senior Member, IEEE.

[10] "Privacy Preserving EHR SystemUsing Attribute-based Infrastructure" Shivaramakrishnan Narayan, Martin Gagné and Reihaneh Safavi-Naini.

[11] "An Enhanced Trust Authorization Based Web Services Access Control Model", R. Joseph Manoj and Dr.A.Chandrasekar, Journal of Theoretical and Applied Information Technology, Vol 64; No 2, pp.522-530 June 2014.

[12] "An Approach to detect and tautology type SQL injection in web services based on web services based on XSchema Validation", R. Joseph Manoj and Dr.A.Chandrasekar and M.D.Anto Praveena, International Journal of Engineering and computer science, Vol 3; No 1, pp.3695- 3699, Jan 2014.