Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 10, October 2015, pg.380 – 387

SURVEY ARTICLE

A SURVEY ON DETECTING DATA DEDUPLICATION AND PROVIDE SECURITY IN CLOUD COMPUTING

[1]V.SUGANYA, [2]J.VISWANATH
PG STUDENT, ASSISTANT PROFESSOR
[1,2]NPR COLLEGE OF ENGINEERING AND TECHNOLOGY, TAMILNADU, INDIA
[1] suganyacse25@gmail.com; [2] viswaj20@gmail.com

Abstract— There is no subsisting result can quietly support the ownership of both exactitude and integrity for the query corollary (result), notably in the case when the fraudulent cloud service provider (CSP) deliberately recompense (returns) an empty set for the query asked by the user.The new verifiable auditing scheme for outsourced database, which can concurrently achieve the exactitude and integrity of search results even if the fraudulent CSP deliberately recompense an empty set. The proposed system achieves the desired security properties even in the encrypted outsourced database. In the Proposed System, the Deduplication method detects and avoids the duplicate data from the data owner.

Keywords— deduplication, cloud service provide (csp), outsourced database.

1. INTRODUCTION

Cloud computing is the distribution of services like software, platform, infrastructure over the Internet. The infrastructure as services provides both hardware and software as a service by virtualization technology to the cloud users. Virtualization is a process of creating a virtual version of os, server, hardware, software. Virtual machine is like computer running within a computer and known as “guest” machine. VMI formats are supported by hypervisor like Xen, Kvm, VMware, Virtual box etc. The large scale VM deployment causes the burden in storage and provisioning the VM and it is overcome by data deduplication techniques. In storage of VMI leads to duplication of files in storage system. Data deduplication eliminates the redundant data in storage system which improves the utilization of storage. In the deduplication process, redundant data is deleted only one copy or single instance of the data to be stored in the database. Cloud computing focuses on maximizing the effectiveness of the shared resources.

Cloud resources are usually not only shared by multiple users but are also dynamically reallocated per demand. This can work for allocating resources to users. The term "moving to cloud" also refers to an organization moving away from a traditional CAPAX model to the OPEX model. The present availability of of high-capacity networks, low-cost computers and storage devices as well as the widespread adaptation of hardware virtualization, service-oriented architecture and automatic and utility computing have led to a growth in cloud computing. Here there are the chances to duplicate the data and store in to the csp we have to avoid duplication of data we introduce a hash algorithm (sha3) providing a sha3 algorithm is used to create unique hash key for each files to avoid duplication the duplication process is checked by third party.

SHA-3 was known as Keccak and is a hash function designed by Guido Bertoni, Joan Daemen, Michaël Peeters, and Gilles Van Assche. MD5 and SHA-0 have been shown to be susceptible to attacks, along with theoretical attacks on SHA-1. NIST thus defined there was a need for a new hashing method which did not use the existing methods for hashing, and setup a competition for competing algorithms. In October 2012, Keccak won the NIST hash function competition, and is proposed as the SHA-3 standard. It should be noted that it is not replacement SHA-2, which is currently a secure methods. Overall Keccak uses the sponge construction where the message blocks are XORed into the initial bits of the state, and then invertibly permuted.

2. LITERATURE REVIEW

2.1 Verifiable Auditing for Outsourced Database in Cloud Computing

The database outsourcing enables the data owner to give the database management to a cloud service provider (CSP) that provides various database services to different users. Here there does not perfectly support the properties of both correctness and completeness for the query results. The csp provide empty set for the users request.to avoid this we propose a new verifiable auditing scheme for outsourced database, which is simultaneously achieve the correctness and completeness of search results even if the fraudulent CSP intentionally returns an empty result. Here we can achieve the security even in the encrypted outsourced database. Further, the proposed system can be protracted to support the dynamic database setting by joining the concept of verifiable database with updates.in checking process we used tuple merkle hash tree which is used to provide signature for each individual attribute to check the correctness and completeness of the data. Bloom filter is used to check whether the data is present or not Evdokimov's scheme is used here to encrypt the data. The advantage is The user can efficiently perform verifiable auditing for the result returned by the CSP. The disadvantage is the uploading and downloading time is high. There is no scheme to check whether the data is already stored or not.so it consumes more space in csp.

2.2 Provable data possession at untrusted stores

The system introduce a model for provable data possession (PDP) that allows a client that has stored data at an untrusted server to verify that the server possesses the original data without retrieving it.The client maintains a constant amount of metadata to verify the proof. The challenge/response provably-secure PDP schemes. advantage is It is used to disguised blocks (called sentinels) hidden among regular file blocks in order to detect data modification by the server the disadvantage is security of the scheme is not proven.

2.3 Network applications of bloom filters: A survey

A Bloom Filter is an ingenious randomized data-structure for concisely representing a set in order to support approximate membership queries. The space efficiency is achieved at the cost of a small probability of false positives. Yet, Bloom's beautiful approach has seen a sudden resurgence in a variety of large-scale network applications such as shared web caches, query routing, and replica location. This survey presents a plethora of recent uses of this old data structure, its modern variants, and the mathematical basis behind them, with the aim of making these ideas available to a wider community and the hope of inspiring new applications. Although Bloom filters were invented in the 1970's and have been heavily used in database applications they have only recently received widespread attention in the networking literature. This survey presents a plethora of recent uses of Bloom filters in a variety of network contexts, with the aim of making these ideas available to a wider community and the hope of inspiring new applications. Bloom filters can be used for summarizing content to aid collaborations in overlay and peer-to-peer networks. Bloom filters allow probabilistic algorithms for locating resources. Packet routing: Bloom filters provide a means to speed up or simplify packet routing protocols. Measurement: Bloom filters provide a useful tool for measurement infrastructures used to create data summaries in routers or other network devices. A Bloom filter is a space-efficient representation of a set or a list that handles membership queries. As the system have seen in this survey, there are various examples where one would like to use a list in a network. Especially when space is an issue, a Bloom filter may be an excellent alternative to keeping an explicit list. The drawback of using a Bloom filter is that it introduces false positives.

It represents a set for membership queries, with false positives. Probability of false positive can be controlled by design parameters. When space efficiency is important, a Bloom filter may be used if the effect false positives can be used if the effect of false positives can be mitigated. Two Bloom filters representing sets S1 and S2 with the same number of bits and using the same hash functions. A Bloom filter that represents the union of S1 and S2 can be obtained by taking the OR of the bit vectors θA Bloom filter can be halved in size. Just OR the first and second halves of the bit vector when hashing to do a lookup, the highest order bit is masked. When a counter does overflow, it may be left at its maximum value. Its maximum value. This can later cause a false negative only if eventually the counter goes down to 0 when it should have remain at nonzero. The expected time to this event is very large but is something the system need to keep in mind for any application that does not allow false negatives. Wherever a list or set is used, and space is at a premium, a Bloom filter maximum, a Bloom filter may be used if the effect of false positives can be mitigated 0 No false negative With a counting Bloom filter, false negatives are possible, albeit highly unlikely.

The effect of a false positive must be carefully considered for each specific application to determine whether the impact of false positives is acceptable. There seems to be a great deal of room to develop variants or extensions of Bloom filters for specific applications. For example, the system have seen that the counting Bloom filter allows for approximate representations of multi-sets, or allows one to track sets that change over time through insertions and deletions. Since Bloom filters have received comparatively little attention from the algorithmic community, there may be a number of improvements to be found. Because of their simplicity and power, the system believe that Bloom filters will continue to applications in networks systems in new and interesting ways. Advantage is it can be used for summarizing content to aid collaborations in overlay and peer-to-peer networks. Bloom filters allow probabilistic algorithms for locating resources. The disadvantage is it cannot perform a deletion by reversing the process.

2.4 Secure outsourcing of scientific computations

This paper investigate the outsourcing of numerical and scientific computations using the following framework: A customer who needs computations done but lacks the computational resources (computing power, appropriate software, or programming expertise) to do these locally would like to use an external agent to perform these computations. This currently arises in many practical situations, including the financial services and petroleum services industries. The outsourcing is secure if it is done without revealing to the external agent either the actual data or the actual answer to the computations. The secure and private outsourcing of linear algebra computations, that enable a client to securely outsource expensive algebraic computations (like the multiplication of large matrices) to a remote server, such that the server learns nothing about the customer's private input or the result of the computation, and any attempted corruption of the answer by the server is detected with high probability. The computational work performed at the client is linear in the size of its input and does not require the client to locally carry out any expensive encryptions of such input. The computational burden on the server is proportional to the time complexity of the current practically used algorithms for solving the algebraic problem (e.g., proportional to n^3 for multiplying two n x n matrices). The improvements we give are: (i) whereas the previous work required more than one remote server and assumed they do not collude, our solution works with a single server (but readily accommodates many, for improved performance); (ii) whereas the previous work required a server to carry out expensive cryptographic computations (e.g., holomorphic encryptions), our solution does not make use of any such expensive cryptographic primitives; and (iii) whereas in previous work collusion by the servers against the client revealed to them the client's inputs, our scheme is resistant to such collusion. As in previous work, we maintain the property that the scheme enables the client to detect any attempt by the server(s) at corruption of the answer, even when the attempt is collusive and coordinated among the servers.

In this paper, the system propose a new secure outsourcing algorithm for (variable-exponent, variable-base) exponentiation modulo a prime in the two untrusted program model. Compared with the state-of-the-art algorithm, the proposed algorithm is superior in both efficiency and check ability. Based on this algorithm, we show how to achieve outsource-secure Cramer-Showup encryptions and Scour signatures. This paper then propose the first efficient outsource-secure algorithm for simultaneous modular exponentiations. Finally, we provide the experimental evaluation that demonstrates the efficiency and effectiveness of the proposed outsourcing algorithms and schemes. The disguise process should be as lightweight as possible, e.g., take time proportional to the size of the input and answer. The disguise pre-processing that the customer performs locally to "hide" the real computation can change the numerical properties of the computation so that numerical stability must be considered as well as security and computational performance This paper show that no single disguise technique is suitable for a broad range of scientific computations but there is an array of disguise techniques available so that almost any scientific computation could be disguised at a reasonable cost and with very high levels of security. These disguise techniques can be embedded in a very high level, easy-to-use system (problem solving environment) that hides their complexity. Advantage is It is easy to use. It can hide their complexity. Disadvantage is it is not clear that there is sufficient demand for such a system to justify this investment.

2.5 Selective and authentic third-party distribution of XML documents

Third-party architectures for data publishing over the Internet today are receiving growing attention, due to their scalability properties and to the ability of efficiently managing large number of subjects and great amount of data. In a third-party architecture, there is a distinction between the Owner and the Publisher of information. The Owner is the producer of information, whereas Publishers are responsible for managing (a portion of) the Owner information and for answering subject queries. A relevant issue in this architecture is how the Owner can ensure a secure and selective publishing of its data, even if the data are managed by a third-party, which can prune some of the nodes of the original document on the basis of subject queries and access control policies. An approach can be that of requiring the Publisher to be trusted with regard to the considered security properties satisfying these requirements in a web environment is very difficult since large systems cannot be easily verified to be secure and are often penetrated. In this paper, we propose a first step towards secure publishing of XML data over the Web by suggesting a scalable architecture that distinguishes between the Owner and the Publisher of information. In this paper, we show how this capability can be accomplished without requiring the Publisher to keep a copy of the access control policies. With a set of digital signatures generated by the Owner and no trust required of the Publisher, we show that a subject is able to verify the authenticity of a query response, and, under specific conditions, the completeness of a query result, with respect to the access control policies stated by the information Owner. The key point of our approach is that even though the system do not require the publisher to be trusted with respected to authenticity and completeness, the system is able to ensure at the same time. This capability is attained using a combination of digital signature and hashing techniques. Our approach requires that the subject first subscribes the owner. During the subscriptions phase the owner determines which access control policies apply to the subject on the basis of the subject credentials.

This approach offers the benefit of enabling a uniform management of XML data and related security information. Completeness verification is a difficult issue and greatly depends on the kinds of query that are submitted to a publisher. In this paper the system present the strategies devised for two common kinds of query. Queries of the first kind are queries whose evaluation against a document d returns a sub tree. However, the serious drawback of this solution is that large Web-based systems cannot be easily verified to be secure and can be easily penetrated. For these reasons, we propose an alternative approach, based on the use of digital signature techniques, which does not require the Publisher to be trusted. The security properties we consider are authenticity and completeness of a query response, where completeness is intended with regard to the access control policies stated by the information Owner. In particular, we show that, by embedding in the query response one digital signature generated by the Owner and some hash values, a subject is able to locally verify the authenticity of a query response. Moreover, we present an approach that, for a wide range of queries, allows a subject to verify the completeness of query results. In this paper, the system has presented an approach for secure web publishing of XML documents. With a set of digital signatures generated by the owner and no trust required of the publisher, the system has shown that a subject can verify the authenticity of a query response. This latter capability is an important feature of our approach since it can serve that is secure against denial-of-service attacks making a distance between the owner and the publisher offers two benefits. First, in any decentralized architecture, such a solution offer the advantage of being scalable and of reducing the risk that the owner becomes the bottleneck of the entire system. Therefore, such a system could be coupled, in our approach, with our publisher component to provide data authenticity and completeness, without the overhead of extensive modifications. Advantage is It used to verify the authenticity of a query response. Disadvantage is Referring a third party for getting appropriate certificate is a defect.

2.6 Verifiable delegation of computation over large datasets

The system study the problem of computing on large datasets that are stored on an untrusted server. The system present the first practical verifiable computation scheme for high degree polynomial functions. Such functions can be used, for example, to make predictions based on polynomials fitted to a large number of sample points in an experiment. In addition to the many non-cryptographic applications of delegating high degree polynomials, The system use our verifiable computation scheme to obtain new solutions for verifiable keyword search, and proofs of irretrievability the polynomials fitted to a large of sample points in an experiment. The advantage is it allows the client to insert and delete values as well as update the value at any cell by sending a single group of element to the server after retrieving the current value stored in the cell the disadvantage is it does not require expensive generation of primes per operation.

3. CONCLUSIONS

In our conclusion we have to avoid the Duplication of Data by Deduplication Method using SHA algorithm and AES algorithm for security and confidence of users the system described the integrity auditing of outsourced database in cloud computing. The system has proposed a new verifiable auditing scheme. This project used Hash Key Algorithm and which can achieve the verifiability of search result even if the result is an empty set. This project used deduplication process supports common database operations such as selection and projection

REFERENCES

[1] Jianfeng Wang, Xiaofeng Chen, Xinyi Huang, Ilsun You, and Yang Xiang,"Verifiable Auditing for Outsourced Database in Cloud Computing,"IEEE Transactions on Computers, Citation information: DOI 10.1109/TC.2015.2401036.

[2] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable data possession at untrusted stores," in Proc. 14th ACM conference on Computer and communications security, pp. 598-609, 2007.

[3] B. Andrei and M. Michael, "Network applications of bloom filters: A survey," Internet Mathematics, vol. 1, no. 4, pp. 485- 509, 2004.

[4] M. J. Atallah, K. N. Pantazopoulos, J. R. Rice, and E. H. Spafford, "Secure outsourcing of scientific computations," Advances in Computers, vol. 54, pp. 215-272, Jan. 2002.

[5] E. Bertino, B. Carminati, E. Ferrari, B. M. Thuraisingham, and A. Gupta, "Selective and authentic third-party distribution of XML documents," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 10, pp. 1263-1278, Oct. 2004.

[6] S. Benabbas, R. Gennaro, and Y. Vahlis, "Verifiable delegation of computation over large datasets," in Proc. 31st Annual Cryptology Conference-CRYPTO 2011, LNCS 6841, Springer, pp. 111-131, 2011.