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. 2, Issue. 4, April 2013, pg.402 – 408

RESEARCH ARTICLE

Modified Security Frame Work for PIR Cloud Computing Environment

J. Sinduja[1], S. Prathiba[2]

[1]Department of Computer Science and Engineering, Bharat University, Chennai, India

[2]Department of Computer Science and Engineering, Bharat University, Chennai, India

Abstract— Computational Private Information Retrieval (cPIR) protocols allow a client to retrieve one bit from a database, without the server inferring any information about the queried bit. These protocols are too costly in practice because they invoke complex arithmetic operations for every bit of the database. Our approach assumes a disk-based architecture that retrieves one page with a single query. Our results indicate that pCloud reduces considerably the query response time compared to the traditional client/server model, and has a very low communication overhead. Additionally, it scales well with an increasing number of peers; achieving a linear speedup Data outsourcing is a new paradigm in which a third party provides storage services. This is more cost effective for the user as there is no need of purchasing expensive hardware and software for data storage. Before data out sourcing can become viable, the data provider needs to guarantee that the data is secure, be able to execute queries on the data, and the results of the queries must also be secure and not visible to the data provider. Data encryption, Homomorphic Encryption, Secret Sharing algorithms and Private Information Retrieval (PIR) are the techniques widely used for secure data outsourcing. CIA (Confidentiality, Integrity and Availability) are the challenging issues associated with data storage management with/without data outsourcing. In this paper the performance of two secret sharing algorithms are compared. The Shamir's secret sharing algorithm and Rabin's Information Dispersal Algorithm (IDA) are implemented in a private cloud setup using the Open Stack Cloud framework.

Key Terms: - Data Security; Cloud; Secret sharing; Information Dispersal

I. INTRODUCTION

Cloud Computing can be defined as the shifting of computing resources like processing power, network and storage resources from desktops and local servers to large data centres hosted by companies like Amazon, Google, Microsoft etc. These resources are provided to a user or company on highly scalable, elastic and pay - as- you-use basis. It reduces the administrative and maintenance cost of IT organizations. From an individual's perspective Cloud Computing is a revolutionary concept as it removes the obstacles created due to lack of finance and resources thus enabling easy large scale deployment an application.

Computing resources provided by cloud vendors can be categorized as computing power, network resources (bandwidth, IP addresses etc.) and storage. The cloud services that were provided earlier can be classified into three categories: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). But, today there is no category, and there is Everything-as-a-Service (Xaa) several free and reliable online storage services available to the users are Apple iCloud, Microsoft SkyDrive, Google Drive, Amazon S3, Dropbox and Box. As the use of these services becomes widespread, security of the outsourced user data becomes an important research topic.

The parameters that are taken into consideration for data security are Confidentiality, Integrity, Availability and Performance. [1]- [3] have emphasized the importance of ensuring remote data integrity. The problem of outsourcing data faces the following obstacles:

1.1. STORAGE OF DATA AT AN UNTRUSTED HOST

Though a service provider gives the guarantee of protecting the privacy of user data, the reality is that the data is physically located in some country and is subject to the local rules and regulations. For example, according to the USA PATRIOT Act the government can access data being hosted by a third party without the permission or knowledge of the user or company using the hosting services [4]. In the recent Megaupload trial regarding the fate of digital files belonging to some 60 million global users there is a possibility that the court will allow Carpathia Hosting, the company that has maintained the servers at its own expense since Megaupload was taken down, to delete the information on them or possibly sell off the servers[5]. Moreover most cloud computing vendors give users little control over where data is stored. Under such circumstances it becomes paramount for a user or company to ensure the confidentiality of his data before it is moved off premise.

1.2. AVAILABILITY OF DATA

Data availability and durability are vital for cloud storage providers, as data loss or unavailability can be damaging to the business. This is usually achieved by replicating the data without the knowledge of the user. Regardless of the precautions taken by a storage service provider we have had major cloud outages recently. VMware Cloud Foundary was down on 25th and 26th April 2011 and Microsoft Azure was out on 28th February 2012[6][7]. Amazon's S3 cloud storage service replicates data across "regions" and "availability zones" so that data and applications can be available even in the face of a disaster affecting an entire location. The user should carefully understand the details of the replication scheme. Though Amazon guarantees that an application using multiple availability zones will not suffer any down time, Amazon web services outage on 21st April 2011 took down several online sites like Reddit, HotSuite, FourSquare and Quora [8]. In view of these events, it is essential for a user or company to ensure the availability of his data without being dependent on the storage service provider.

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 and discusses the System Model and Security framework. Section 4 and section 5 describe the two widely used algorithms for secure data storage. The experimental setup is presented in section 6 and the conclusion in section 7.

II. RELATED WORK

Data encryption, Homomorphic Encryption, Secret Sharing algorithms and Private Information Retrieval (PIR) are the techniques widely used for secure data outsourcing. This section provides a review of the various works done in the above mentioned techniques.

Data encryption has traditionally been used to provide confidentiality while outsourcing data to the service providers. Hacigumus et al. [9] discusses a method for executing queries over encrypted data, at the service provider's site, and suggests splitting a query into two parts, namely the server query and client query. The server query is executed over the encrypted data at the service provider and the other part over the results of server query, at the client side. Hore et al. [10] describes techniques for building privacy preserving indices on sensitive attributes of a relational table, and provides an efficient solution for data bucketization. Agrawal et al. [11] highlights the benefits of using the Order Preserving Encryption Scheme (OPES) for querying numeric data. Yang et al. [12] shows the importance of building secure indices and demonstrate how they help to achieve reduction in the query execution time. Sion [13] emphasizes that a system that employs encryption for data outsourcing is secure if it ensures *correctness, confidentiality and data access privacy.* He also discusses about how these components are inter-related.

Outsourcing by encrypting the whole data is prohibitive in terms of performance and it does not solve the problem of availability. Moreover encryption techniques are secure based on the present computing power but with the advances being made in computing speeds this may not be always true. For example, distributed.net and Electronic Frontier Foundation joined hands in January 1999, to break a DES key, which was previously thought to be unbreakable, in 22 hours and 15 minutes. Additionally, secure maintenance of keys used for encryption becomes important [14].

Homomorphic Encryption performs "computation on the ciphertext before decrypting it first" [15]. Scientists predict that this promising technique would be of immense help in implementing many future cloud computing applications securely [14]. Ge et al. [16] discuss about computing efficient and accurate results for the SUM and

AVG queries using a secure, modern homomorphic scheme. Homomorphic encryption is computationally expensive and due to its dependence on the public key cryptosystem, it is difficult to implement. [17].

The concept of Private Information Retrieval (PIR) was first discussed in [18]. Private information retrieval protocols intend to hide the queries performed by the user on a public database, stored on a set of servers. By providing the privacy of user queries the PIR protocols tend to hide the user's intensions from the Service provider. The idea of PIR has been extended to Symmetric Private Information Retrieval (SPIR), in which the privacy of user data is the main concern. In [19], Sion and Carburnar have extensively discussed about the practical infeasibility of implementing the single-server computational PIR protocol. Williams and Sion [20], describe a new PIR technique developed using the Oblivious RAM (ORAM) technique. They illustrate the much lesser communication latency and computational complexity the technique exhibits, with a small increase in the space complexity.

Secret Sharing techniques are also one of the widely used techniques for data outsourcing. Two popular secret sharing techniques are the Shamir's Secret Sharing method and Rabin's Information Dispersal Algorithm (IDA). In Shamir's secret sharing [21] algorithm, a file F to be outsourced is split into n parts F_1, F_2, F_3….F_n, such that each file F_i, $i \leq n$, is padded with redundant information to make its size same as that of F. The file F can be retrieved if k out of n pieces is available. Shamir calls this as *threshold(k,n)*. Shamir's secret sharing method is information theoretically secure.

Rather, Rabin [22] suggested splitting a secret S into n pieces such that a person can obtain the secret only if $k < n$ of these pieces are available, where k is the threshold. Here, each secret S_i, $i \leq n$, is of size $|S|/k$, where $|S|$ is the size of the secret. The total sizes of all the secrets are $(n/k)*|S|$.

Thus, with the Rabin's Information Dispersal Algorithm, the storage complexity is reduced. But, the security flaw in this method is that, if the data exhibits some pattern frequently, and that the attacker gets hold of $m < k$ slices, then there are great possibilities for him to get the secret S [23].

III. SYSTEM MODEL AND FRAMEWORK

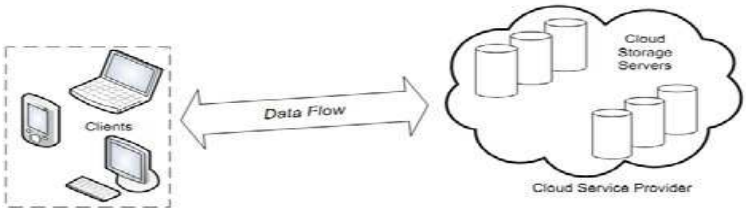

Fig-1. Cloud data storage architecture

The cloud data storage architecture used in this work is based on the model proposed by Cong Wang et. al [24], as shown in Fig-1. The different entities are the Client and Cloud Storage Server.

Client: The end user who has large amount of data to store in the cloud and relies on the service provider for maintenance. This can either be an individual user or a large organization.

Cloud Storage Server: An entity, which is managed by a Cloud Service Provider, has significant storage space and computation resource to maintain client's data.

3.1 SECURITY FRAMEWORK

Fig-2 shows the framework for querying of confidential data stored in a cloud environment. There are two major types of servers in cloud computing which are storage server and computational server. Storage server provides the service related to data storage or updation.

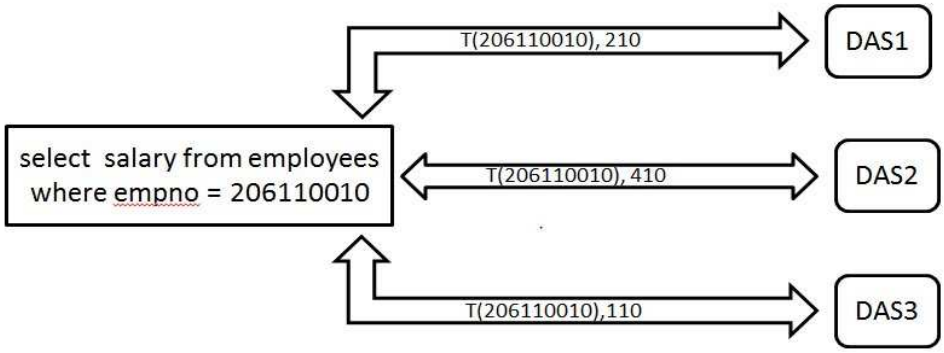

Fig-2 Querying of the secured data

IV. SHAMIR'S SECRET SHARING

Shamir's algorithm is implemented based on the secure order preserving technique discussed in [25]. Shamir's algorithm is applied to each field of the table. The *n* shares obtained are then distributed to different data centers out of which only *k* of the shares are required to achieve the original values. It is possible to support structured querying of the data by parsing the query, extracting the conditional values, transforming the values and appending them back to the original query before it is sent to the server. Inverse of Shamir's Secret sharing algorithm has to be applied to the received data set to get back the intended result. Assume that the table 'employees (emp_no, salary, empname)' is outsourced. The client should be able to execute the following type of queries without revealing the data to any of the Database service providers.

- Exact match queries
- Range queries
- Aggregate queries such as MIN/MAX, MEDIAN, SUM and AVERAGE

V. RABIN'S INFORMATION DISPERSAL ALGORITHM

Rabin's IDA is implemented using the technique discussed in [17]. Considering k as the threshold value, n as the number of slices, and D as the Data Matrix of size *k x t*, The data to be stored is arranged in terms of the data matrix D and C as the secret Vandermonde Matrix of size *n x k*. The matrix M of size *n x t* is computed as

$$M = C * D \qquad (1)$$

Each of the *n* rows of *M* represents a slice. This modified data is stored at multiple data centers such that none of them have access to $s < k-1$ slices.

Data retrieval can be achieved by obtaining any *k* of the *n* slices and applying the reverse IDA. Consider *M'* to be the *k x t* matrix formed by obtaining the *k slices* of data stored in the cloud and *C'* to be the *k x k* matrix obtained by selecting the corresponding rows of C. Then the data matrix D can be retrieved as:

$$D = C'^{-1} * M' \qquad (2)$$

Even with the loss of *(n-k)* slices, the data can be reproduced thus ensuring availability. A message authentication code can be applied to the test data before dispersal to achieve integrity.

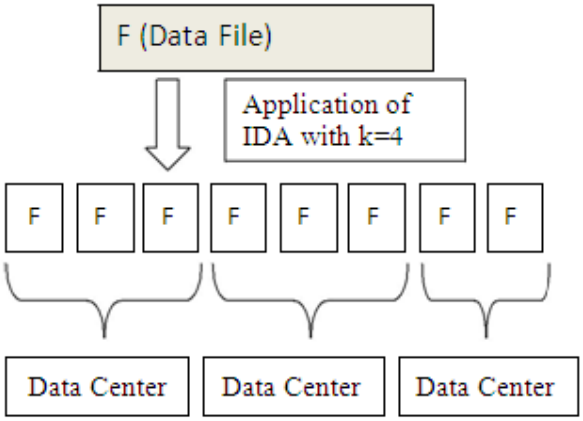

Fig-3. Information Dispersal

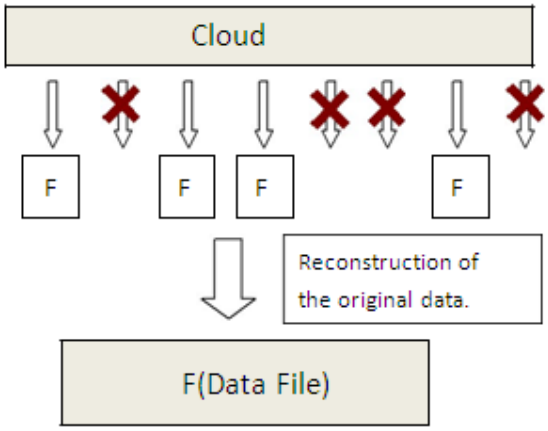

Fig- 4. Information Recovery

VI. EXPERIMENTAL SETUP

The OpenStack cloud framework was used to compare the performance of the two methods. OpenStack is an IaaS cloud computing project. Openstack is gaining importance in the academia as well as in the industry. Major cloud players like AMD, Intel, HP, Linux and Cisco have joined the OpenStack project [26].

Linux machines were used as compute nodes, controller and client. The controller node consisted of the nova-api, network controller, scheduler and the RabbitMQ asynchronous messaging server. The compute nodes had the compute controller component installed in them. The virtual machines were instantiated and run in these compute nodes. Six virtual machines were instantiated, each one representing a cloud storage server. The FlatManager network configuration was used for the setup of the OpenStack cloud.

The Employees sample database developed by Patrick Crews and Giuseppe Maxia was used as the test data. This large database has six separate tables and a total of four million records [27].

The Shamir's Secret Sharing Algorithm was applied to this data with no of shares $n= 3$ and the threshold value $k= 2$. Each of these shares were stored on a separate virtual machine instances (Data Center). It was demonstrated that the data available to each of the Data Center was incomprehensible (Confidentiality). The

client was able to execute all the common SQL queries on this data. It was able to verify the veracity of the data and execute queries even with the outage of *(n-k-1)* data centers, thus ensuring integrity and availability. The drawback of Shamir's approach in a pay-per-use cloud computing model is that the amount of storage required is increased by *n* times.

Rabin's IDA *(n,k)* was applied to the data file at the client side. These files were randomly placed at the storage servers using the SCP (secure copy) protocol in such a way that no server had *k* files. In the implementation of Rabin's IDA, we are successfully able to reconstruct the whole data even when *(n-k)* slices of data were unavailable.

Fig-5 represents the time taken to split a 94 MB into 10 pieces for different threshold values and Fig-6 represents the time taken to combine them. Overhead is calculated as *((combined size of split files/original file size) * 100)* for different values of n and k.

It was observed that there was considerable decrease in the Dispersal time as the overhead decreases. The Recovery time remains almost constant with a maximum variation of 0.4 seconds with the best recovery time at the threshold value of eight. The overhead at this point was 25%

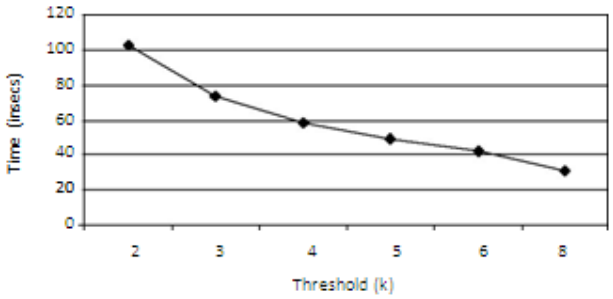

Fig- 5. Information Dispersal for 94 MB file with n=10

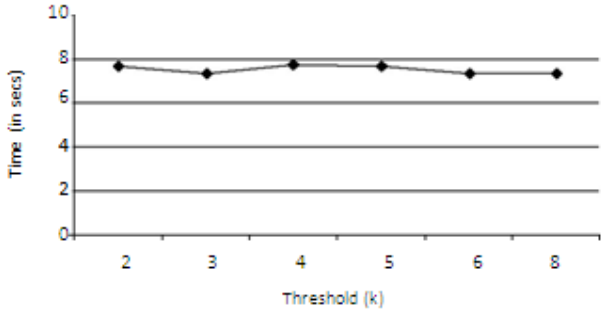

Fig- 6. Information Recovery for 94 MB file with n=10

Using the Shamir's Secret Sharing and the Rabin's IDA, we found that with n=3 and k=2, we were able to transform and distribute the employee sample database to three virtual machines within 47 Seconds and 30 seconds respectively.(justification).

VII. CONCLUSION

Data outsourcing using Rabin's IDA and Shamir's secret sharing methods are compared using the OpenStack cloud framework. The Secret Sharing methods are computationally inexpensive when compared with the traditional encryption techniques.

The security of these methods is not bounded by the computational capabilities of the present hardware. Further secure storage of the encryption keys is done away with.

Taking the above points into consideration we conclude that, owing to the distributed nature of the cloud, information dispersal of data is the more optimal approach for data outsourcing.

REFERENCES

[1] Juels, A. and Burton, J., Kaliski, S.: PORs: Proofs of Retrievability for Large Files. In: Proc. of CCS '07, pp. 584–597 (2007)

[2] Shacham, H., and Waters, B.: Compact Proofs of Retrievability. In: Proc. of Asiacrypt '08, Dec. (2008)

[3] Bowers, K.D., Juels, A., and Oprea, A.: Proofs of Retrievability: Theory and Implementation. In: Cryptology ePrint Archive, Report 2008/175 (2008), http://eprint.iacr.org/

[4] Wikipedia cited 2012: Patriot ACT. [Available online at http://en.wikipedia.org/wiki/Patriot_Act]

[5] David Saleh Rauf, Politico, cited 2012: Megaupload data: Hosting company chafes at maintenance. [Available online at http://www.politico.com/news/stories/0312/74354.html]

[6] Dekel Tanke, Cloud Foundary, cited 201l: Analysis of April 25 and 26, 2011 Downtime. [Available online at http://support.cloudfoundry.com/entries/20067876-analysis-of-april-25-and-26-2011-downtime]

[7] Bill Laing, Windows Azure, cited 2012: Summary of Windows Azure Service Disruption on Feb 29th, 2012. [Available online at http://blogs.msdn.com/b/windowsazure/archive/2012/03/09/ summary- of-windows-azure-service-disruption- on-feb-29th-2012.aspx]

[8] Amazon Web Services team, cited 2011: Summary of the Amazon EC2 and Amazon RDS Service Disruption in the US East Region. [Available online at http://aws.amazon.com/message/65648/]

[9] H. Hacigumus, B. R. Iyer, C. Li, and S. Mehrotra, "Executing SQL over encrypted data in the database service provider model," in Proc of the ACM SIGMOD Conf., 2002.

[10] B. Hore, S. Mehrotra, and G. Tsudik, "A privacy-preserving index for range queries," in Proc. of the VLDB Conf., 2004, pp. 720–731.

[11] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in Proc. of the ACM SIGMOD Conf., 2004, pp. 563–574.

[12] Z. Yang, S. Zhong, and R. Wright, "Privacy-preserving queries on encrypted data," in Proc. of the 11 European Symposium on Research In Computer Security, 2006

[13] Sion, R.: Secure data outsourcing. In: Proc. of the VLDB Conf., pp. 1431–1432 (2007)

[14] Data Encryption Standard [Available online at http://pustakalaya.org/wiki/wp/d/Data_Encryption_Standard.htm]

[15] Craig Stuntz, What is Homomorphic Encryption, and Why Should I Care? [Available online at http://blogs.teamb.com/craigstuntz/2010/03/18/38566/]

[16] Ge, T., and Zdonik, S.B.: Answering aggregation queries in a secure system model. In: Proc. of the VLDB Conf., pp. 519–530 (2007)

[17] S. Wang, D. Agrawal, A.E. Abbadi: A Comprehensive Framework for Secure Query Processing on Relational Data in the Cloud. Secure Data Management 2011: 52-69

[18] Chor, B., Goldreich, O., Kushilevitz, E., and Sudan, M.: Private information retrieval In: Journal of the ACM, vol. 45, no. 6, pp. 965–982 (1998)

[19] Sion, R., and Carbunar, B.: On the computational practicality of private information retrieval. In: Proc. of the Networks and Distributed Systems Security (2007)

[20] P. Williams and R. Sion, Usable PIR. NDSS, 2008.

[21] Shamir, A.: How to share a secret. In: Commun. ACM, vol. 22, no. 11, pp. 612–613 (1979)

[22] Rabin, M.O.: Efficient dispersal of information for security, load balancing, and fault tolerance. In: Journal of The ACM 36(2), pp. 335–348 (1989)

[23] Resch, Jason; Plank, James (February 15, 2011). "AONT-RS: Blending Security and Performance in Dispersed Storage Systems". Usenix FAST'11, 2011

[24] Cong Wang, Qian Wang,. Kui Ren, Wenjing Lou,: Ensuring data storage security in Cloud Computing, Quality of Service, 2009. IWQoS. 17th International Workshop on , vol. 186, pp.1-9, (2009)

[25] D. Agrawal, A.E.Abbadi, F.Emekci, A. Metwally: Database Management as a Service: Challenges and Opportunities, IEEE International Conference on Data Engineering, 2009

[26] http://www.openstack.org

[27] https://launchpad.net/test-db/

