Data Provenance for Cloud Computing using Watermark

Muhammad Umer Sarwar*, Muhammad Kashif Hanif†, Ramzan Talib‡ Bilal Sarwar§, and Waqar Hussain¶
Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—The term data is new oil which has become a proverb due to large amount of data generation from various sources. Processing and storing such tremendous amount of data is beyond the capabilities of traditional computing system. Cloud computing preferably considered next-generation architecture due to dynamic resource pools, low cost, reliability, virtualization, and high availability. In cloud computing, one important issue is to track and record the origin of data objects which is known as data provenance. Major challenges to provenance management in distributed environment are privacy and security. This paper presents data provenance management for cloud computing using watermarking technique. The experiment is performed by using both visible and hidden watermarks on shared data objects stored in cloud computing environment. The experimental results demonstrate the efficiency and reliability of proposed technique.**

***Keywords*—Cloud computing; data provenance; watermark; security; visible watermark; invisible watermark**

I. Introduction

During the last 20 years, continuous development in web technology has produced a huge collection of data. Before 2020, people and linked data objects will approximately generate the billions gigabytes of data that will have an influence on daily life [1]. It would be difficult to manage such a huge volume of data with traditional processors. Organizations have started to manage such large amount of data by shifting their services over the cloud. According to *Gartner's 2011 CIO Agenda Survey*, most of organizations will depend on the cloud computing more than half of their IT Services before 2020. This prototype shifting reduced the costs to manage the software and hardware resources.

Cloud computing provide accessibility of data, files, programs, and services from web browser over internet. Cloud computing stores the software, programs, and other computing applications to a central location. The management of resources on cloud have security challenges. One of the important issue is to ensure data integrity. It is dangerous to guarantee data integrity in cloud computing for results assembling process [2]. One possible solution to ensure data security is *Data Provenance*. In cloud computing, the term data provenance is defined as the original source of shared data objects.

In this paper, a watermarking technique is used to store provenance information of shared data objects in cloud computing. This technique will help to find the original source of data objects in cloud computing. The experiment is performed by implementing the presented watermarking technique in an open source platform known as *ownCloud* [3]. This approach is used to answer the following questions related to data security in cloud computing: What is the original source of data object? Who is the owner of data object? How much reliable is the original source? Who modified the data object? Finally, the efficiency and reliability of proposed technique is measured.

The rest of the paper is structured in different sections. Section II presents related work. Section III describes cloud computing along with its architecture and security challenges. In Section IV, data provenance and techniques for maintaining provenance are presented. Proposed methodology is presented in Section V. Section VI discusses the results. Finally, we conclude the work with outcomes in Section VII.

II. Related Work

The rapid and large increase in data poses the problems of data security in cloud computing. Numerous watermarking techniques have been proposed by the researchers for ensuring the security of shared data. These techniques are categorized on the basis of types of watermark like digital watermark, visible watermark, invisible watermark, and cover type [4]. These techniques can also be characterized by data provenance, data lineage, usability, and robustness [4].

A few watermarking techniques have been used in order to ensure the data security and integrity [5], [6]. Some researchers have used watermarking techniques for data provenance that includes fragile [7] and novel [8] watermarking. In these techniques, some important issues like data security, usability, robustness, distortion, and capacity are taken under consideration. However, these techniques lacks to solve the data provenance problem to optimum level. Sarwar et al. presented a package watermarking technique to ensure data provenance in database systems [9].

Tiwari and Sharma studied various semi fragile watermarking algorithms by using different methods like image quality matrices, insertion and verification methods [10]. Zhang et al. proposed a gray scale watermark pre-processing technique. Their work provides robustness of video watermarking for copyright protection. This approach almost maintains the same bit rate and also provides good visual quality [11]. Some researchers have worked on the problem of data provenance [12], [13]. However, their work only emphasis on the semantic analysis of data provenance information.

III. Cloud Computing

Cloud computing is relatively a new computing model which provides high performance computational services at a minimal cost. Some famous organizations in the field of

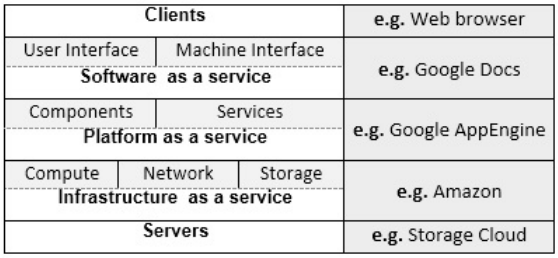

Fig. 1. Cloud computing architecture.

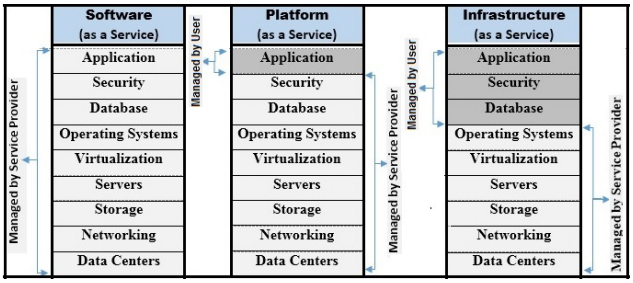

Fig. 2. Cloud computing services (CIO Research Center, 2010) [19].

IT such as Google, Microsoft, and Amazon have shifted their cloud services over the Internet [14]. Cloud computing is called the fifth utility in the line of electricity, water, telephone and gas [15]. It provides the facility of storing and accessing files from any where in the world based on access permissions. Main advantage of cloud computing is to allow the users to carry out their every day computing operations using cloud computing [16]. The term cloud computing is defined by the researchers in various ways. Buyya, et al. [17] defined cloud computing is distributed computer system consisting of the collection of interconnected and virtualized computers that are supplied with strength and can be obtained as one or more conventional Service Level Agreements (SLA) based processing resources through negotiation between service providers and consumers. According to Vaquero, et al. [18], clouds are a large amount of virtualized resources (such as hardware, development platforms, and services) that can be accessed in a functional and accessible manner. These resources can be dynamically reconfigured to fit an adjustable load scale enabling an optimum consumption of resource. This resource pool is abused by a pre-use payment model where the warranty is offered by the infrastructure provider through custom service level agreements.

Cloud computing can be classified into three well-known and frequently used service models. These models are known as Software as a Service (SaaS), Platform as Service (PaaS), and Infrastructure as Service (IaaS). Fig. 1 shows hierarchy of these services with examples. These service model provides specific features and functionalities. The major difference between these service models is depicted in Fig. 2. SaaS is famous and well-known type of cloud service for common users. SaaS applications are hosted on remote server and are managed by the service provider. These applications are accessible to users through a web browser on Internet. Some well known SaaS applications are Google Apps and DropBox. PaaS is similar to SaaS in many ways. Instead of delivering the applications over the web, PaaS provides an environment for users to create their own applications. However, rest of the computing elements except applications are managed by the PaaS service provider as shown in Fig. 2. Example of PaaS application is Google App Engine. IaaS is the most flexible cloud computing model and allow users to run any applications. These applications can be run on the hardware of IaaS service provider. IaaS offers web-based access to computing power and storage. In this service model, the user does not need to control PaaS infrastructure or SaaS services. Famous applications of IaaS includes Google Engine, Amazon Web Services, and Microsoft Azure.

The cloud computing is categorized in public, private, and hybrid clouds. Private clouds are organized within premises and accessible only to a single organization. Public clouds are deployed off premises and accessible by any user within or outside the organization. Hybrid cloud may be either internal or external. Hybrid cloud contains the characteristics of both private and public clouds [14]. Fig. 3 Illustrates these cloud deployment types [20].

Fig. 3. Cloud deployment models [20].

Although cloud computing is emerging rapidly but at the same time there are severe security challenges. Some security issues have reduced the adoption of cloud computing among its users. One important problem in cloud computing is to ensure the trustworthiness of data object. Thus, increase in demand of cloud computing has raised a few security problems for both of its users and service providers. How the users can ensure the unique ownership of their uploaded data object in the cloud? Every user wishes to know about the availability of their data objects in the cloud [21].

IV. Data Provenance

With the huge growth of data, finding origin and transformation of data becoming an important and challenging task. In many applications like cloud computing, database, and social media network, it is a challenging task to find out the origin of data object. The original source of shared data object in cloud computing is very important in order to take any decision. The term "data provenance" means a procedure to trace and record the origin of data products. The importance of the data source is increasingly recognized by both users and publishers of that data product for a long time. The original source of shared data objects in cloud can be used by scientists or scholars to determine the real ownership who is working on these data products. Likewise, medical research requires severe product quality data checks because errors can harm people's health.

The term data provenance is defined as history of ownership of a shared data object. Provenance can also help to find out the authenticity of a shared data object in cloud computing. In other words, provenance is a term used in

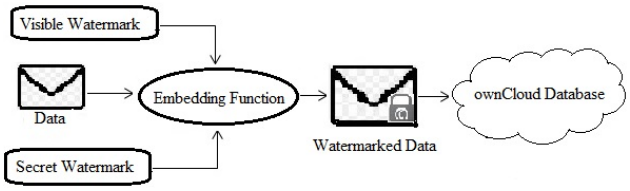

Fig. 4. Watermark embedding process.

diverse areas which describes the history of an object since its creation. In cloud computing, data is shared widely and anonymously. Therefore, the source of data object is required to verify the authenticity of that data product. Cloud users face severe security challenges from both inside and outside the cloud. In fact they have to face some threats from cloud service providers. If the source is provided in cloud computing, users can gain more control over their data. Additionally, cloud users can detect what went wrong with respect to data under its control. In particular, cloud users may be able to verify that nothing has gone wrong with its data products.

Currently, there is no way to trust on the data provenance information in the cloud computing. However, numerous novel techniques are designed and implemented in cloud computing. Some issues of data security and governance in cloud computing are discussed in [22]. A data security framework for cloud computing networks is proposed [23]. Younis and Kifayat provides a survey on secure cloud computing for critical infrastructure [24]. Chen and Zhao [25] analyzed privacy and data security issues in the cloud computing by focusing on privacy protection, data separation, and cloud security. According to a survey [12], characteristics of nine different data provenance techniques are summarized in [12].

V. Methodology

Data security is one of the biggest concerns about cloud computing. Numerous different techniques are introduced by the computing researchers for data protection as already discussed in Section IV. However, there are still some gaps in those techniques which require enhancement. In this section, the main focus is to explain how watermarking technique can be used to maintain original ownership about data in cloud computing. The proposed watermarking technique for data provenance is practically examined using a free and open-source cloud software known as ownCloud. ownCloud is a file sharing server that permits its users to store data objects in a centralized location, much like Dropbox [26]. These data objects can be any type of images or text documents.

In this paper, two important watermarking techniques are used to secure the shared data objects in cloud computing. First technique is to embed the visible watermark which can be seen by everybody who is seeing the data object. This type of watermark ensures trustworthiness of data packets in the cloud. Second technique is to insert the hidden watermark which provides backup facility in case when visible watermark fails to prove trustworthiness of data. Both visible and hidden watermarks are embedded in the data objects when these data packets are created or added for the first time in the cloud. The process of embedding both visible and hidden watermarks in the host data object consists of two steps as shown in Fig. 4. In first step, both types of watermarks are embedded in the data object. Then, in second phase, watermarked data objects are stored in the database of relevant cloud like ownCloud. The experiment is performed with six different images using an open source free cloud known as ownCloud. These images are taken randomly and then both visible and invisible watermarks are embedded in them according to proposed formula which is described in next paragraphs.

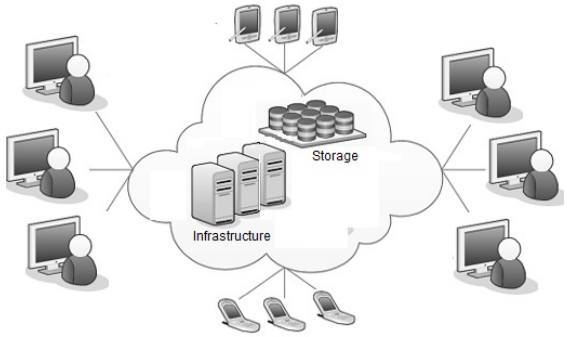

Fig. 5. Cloud computing network.

ownCloud inherits the characteristics of IaaS, a well-known service model of cloud computing as already discussed in Section III. Fig. 5 represents a network diagram between ownCloud and its users. Actually ownCloud infrastructure is installed at a centralized server computer and configured with a static IP address. Then, different kinds of nodes or clients like desktop computers, mobile phones, and laptops are deployed in a network. These clients can access the ownCloud interface through an IP address. All clients can upload and access their data to ownCloud through a web enabled-interface.

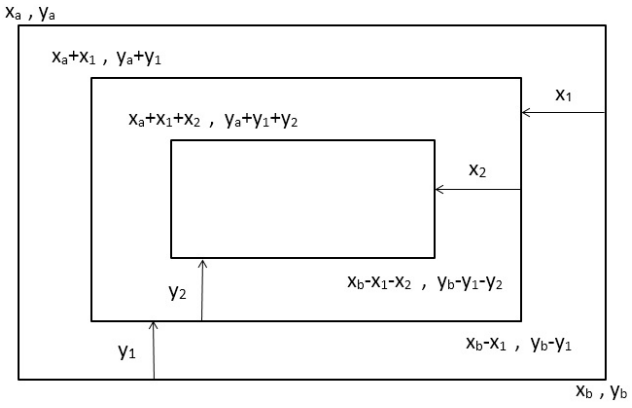

Fig. 6. Watermark placement in images.

Fig. 6 shows the strategy that is adopted to embed both visible and hidden watermarks on the uploaded data object like images. In order to perform the experiment, the whole image is divided into three rectangles. The hidden watermark in inserted into the innermost rectangle while visible watermark is placed into the outer rectangle adjacent to the outermost boundary of the image.

$$x_i = (x_{i-1})/(i \times 10) \quad (1)$$

$$y_i = (y_{i-1})/(i \times 10) \tag{2}$$

Where, $i \geq 1$, x_0, and y_0 is width and height of original image, respectively.

$$\sum_{i=1}^{l} x_i, \sum_{i=1}^{l} y_i \tag{3}$$

$$x_b - \sum_{i=1}^{l} x_i, y_b - \sum_{i=1}^{l} y_i \tag{4}$$

Where, $l \geq 1$

In (1) and (2), the gap between these rectangles in an image is calculated which separates the inner rectangle from outer one. Equation (1) calculates the width of inner rectangle which is reduced by ten percent w.r.t the width of outer rectangle, whereas (2) calculates height of inner rectangle. After calculating the gap from outer most layers, coordinates of first and last point of inner rectangle are calculated in (3) and (4). In this way, the whole area of the image is partitioned into three rectangles and both visible and invisible watermarks are embedded in outer and inner rectangle, respectively.

TABLE I. IMAGE UPLOADING TIME IN CLOUD (SEC)

Sr. No	without watermark	with watermark
1	0	0.27
2	0	0.31
3	0	0.33
4	0	3.57
5	0	5.98
6	0	7.65

VI. RESULT AND DISCUSSION

In this paper, the adopted methodology exposes an interesting open research question on data provenance in cloud computing. Data is increasing in huge amount, it is essential for provenance information to be shared along with the data object. In this work, a watermarking technique is used to store the provenance information about shared data objects in cloud computing. Data objects which are created by the cloud users are stored in the database of ownCloud in two steps. In first step, data is embedded with both visible and hidden watermark and in second step, watermarked data finally stores in ownCloud database.

In this section, a criteria is introduced to evaluate the efficiency and reliability of the adopted methodology. This criteria takes decision upon the results that are generated from the proposed watermarking technique. The efficiency of the proposed watermarking technique is measured in terms of time required to upload a data object from client's local machine to cloud database. For one data object, it's time to watermark the input data object and then to store that watermarked object in ownCloud database. In Table I, it can be seen that for all six different images, time required to upload the watermarked image is greater than that of without watermark image. Another aspect is the size of the uploaded image which is increased when both visible and invisible watermarks are embedded in it.

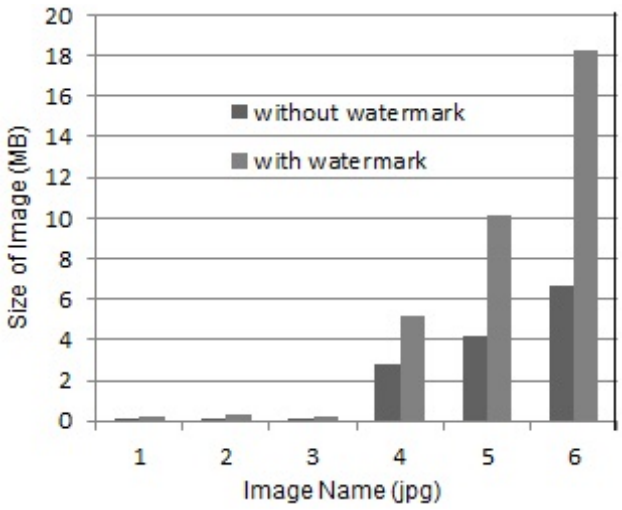

Fig. 7. Size of original and watermarked images.

Size difference of all six images with and without watermarks is depicted graphically in Fig. 7. On the other hand, reliability depends on the nature of watermarks that are embedded on the shared data object in cloud computing. In this scenario, two dissimilar watermarks are used i.e. visible watermark is different from hidden watermark in every aspect. This feature enhances security of data objects in cloud computing.

VII. CONCLUSION

The main focus throughout this paper is on the problem of determining the trustworthiness of data in cloud computing. In order to trust on the cloud data, there is a need to track the origin of data object. To address this problem, a watermarking technique is proposed which stores the information about the origin of data product. This technique uses two important types of watermarks that are visible watermark and hidden watermark. By adopting this methodology, shared data object in the cloud can be safe from the malicious attack that may change or lose the real ownership of that data object. Finally, the efficiency and reliability of this adopted approach is evaluated by calculating the time required to embed both visible and hidden watermarks on data objects. In this paper, the problem of data provenance is focused within a same cloud computing environment. In the near future, some other techniques are expected to ensure the trustworthiness of shared data objects among different cloud computing environments simultaneously.

REFERENCES

[1] S. Sakr, A. Liu, D. M. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 3, pp. 311–336, 2011.

[2] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu, "An approach to evaluate data trustworthiness based on data provenance," in *Workshop on Secure Data Management*. Springer, 2008, pp. 82–98.

[3] A. Patawari, *Getting started with ownCloud*. Packt Publishing Ltd, 2013.

[4] R. Halder, S. Pal, and A. Cortesi, "Watermarking techniques for relational databases: Survey, classification and comparison." *J. UCS*, vol. 16, no. 21, pp. 3164–3190, 2010.

[5] K. Ramya, R. C. Devi, M. Revathi, and P. Annapandi, "Sensor data hiding based on image watermarking using interpolation technique over inter-packet delays." *Applied Mechanics & Materials*, no. 573, 2014.

[6] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1064–1087, 1998.

[7] M. Schäler, S. Schulze, R. Merkel, G. Saake, and J. Dittmann, "Reliable provenance information for multimedia data using invertible fragile watermarks," in *British National Conference on Databases*. Springer, 2011, pp. 3–17.

[8] L. Zhang, Y. Zhu, and L.-M. Po, "A novel watermarking scheme with compensation in bit-stream domain for h. 264/avc," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 1758–1761.

[9] M. U. Sarwar, M. K. Hanif, R. Talib, and M. A. Abbas, "Ensuring data provenance with package watermarking," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, pp. 498–501, 2017.

[10] A. Tiwari and M. Sharma, "Semifragile watermarking schemes for image authentication-a survey," *International Journal of Computer Network and Information Security*, vol. 4, no. 2, pp. 43–49, 2012.

[11] J. Zhang, A. T. Ho, G. Qiu, and P. Marziliano, "Robust video watermarking of h. 264/avc," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 54, no. 2, pp. 205–209, 2007.

[12] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance techniques," *Computer Science Department, Indiana University, Bloomington IN*, vol. 47405, 2005.

[13] M. Greenwood, C. Goble, R. D. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, and T. Oinn, "Provenance of e-science experiments-experience from bioinformatics," in *Proceedings of UK e-Science All Hands Meeting 2003*, 2003, pp. 223–226.

[14] Y. Sun, J. Zhang, Y. Xiong, and G. Zhu, "Data security and privacy in cloud computing," *International Journal of Distributed Sensor Networks*, vol. 2014, 2014.

[15] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599–616, 2009.

[16] W. Voorsluys, J. Broberg, and R. Buyya, "Introduction to cloud computing," *Cloud computing: Principles and paradigms*, pp. 1–44, 2011.

[17] B. Rajkumar, C. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms," *Future Generation Computer Systems. Elsevier Press, Inc*, 2009.

[18] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, pp. 50–55, 2008.

[19] R. K. Perrons, "How the energy sector could get it wrong with cloud computing," *Energy Exploration & Exploitation*, vol. 33, no. 2, pp. 217–226, 2015.

[20] V. Zapolskas, "Securing cloud storage service," 2012.

[21] F. B. Shaikh and S. Haider, "Security threats in cloud computing," pp. 214–219, 2011.

[22] Z. Mahmood, "Data location and security issues in cloud computing," in *International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*. IEEE, 2011, pp. 49–54.

[23] A. Pandey, R. Tugnayat, and A. Tiwari, "Data security framework for cloud computing networks," *International Journal of Computer Engineering & Technology*, vol. 4, no. 1, pp. 178–181, 2013.

[24] M. Younis and K. Kifayat, "Secure cloud computing for critical infrastructure: A survey," *Liverpool John Moores University, United Kingdom, Tech. Rep*, 2013.

[25] D. Chen and H. Zhao, "Data security and privacy protection issues in cloud computing," in *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*, vol. 1. IEEE, 2012, pp. 647–651.

[26] B. Martini and K.-K. R. Choo, "Cloud storage forensics: owncloud as a case study," *Digital Investigation*, vol. 10, no. 4, pp. 287–299, 2013.