A digital forensic model for providing better data provenance in the cloud

Philip M. Trenwith
Software Department
GEW Technologies
Pretoria, South Africa
Email: ptrenwith@gmail.com

Hein S. Venter
Department of Computer Science
University of Pretoria
Pretoria, South Africa
Email: hventer@cs.up.ac.za

***Abstract*—The cloud has made digital forensic investigations exceedingly difficult due to the fact that data may be spread over an ever-changing set of hosts and data centres. The normal search and seizure approach that digital forensic investigators tend to follow does not scale well in the cloud because it is difficult to identify the physical devices that data resides on. In addition, the location of these devices is often unknown or unreachable. A solution to identifying the physical device can be found in data provenance. Similar to the tags included in an email header, indicating where the email originated, a tag added to data, as it is passed on by nodes in the cloud, identifies where the data came from. If such a trace can be provided for data in the cloud it may ease the investigating process by indicating where the data can be found. In this research the authors propose a model that aims to identify the physical location of data, both where it originated and where it has been as it passes through the cloud. This is done through the use of data provenance. The data provenance records will provide digital investigators with a clear record of where the data has been and where it can be found in the cloud.**

Keywords. Digital Forensics, Digital Forensic Investigation, Cloud Computing, data provenance, bilinear pairing technique, chain of custody, annotations

I. Introduction

The goal of this research paper is to investigate how the history of digital objects known as data provenance, can be used in the cloud in order to provide a trace of the path the data has travelled in the cloud.

The traditional model of a Digital Forensic Investigation(DFI) follows a search and seizure approach. This does not scale well in the cloud due to the large set of hosts and data centres that data in the cloud is spread over. Therefore it is necessary to investigate the short comings of the current model and determine the requirements for a model for digital forensic investigations that better suit the architecture of cloud computing environments.

Edmond Locard defines the principle of exchange stating, "whenever two objects come into contact with each other, each object is left with a trace of the other object"[1]. These traces left behind on a digital object as it comes in contact with nodes in the cloud can first of all help identify where the object has been and secondly where it can be found in the cloud. Actively keeping record of trace evidence by appending an identifying tag to the objects metadata as the object is transmitted through the cloud will identify where the object has been. This is similar to a person's passport that is stamped each time the person enters a country. This practise address a specific challenge faced by investigators: identifying the physical location of data in the cloud [2]. This approach will help to provide the digital forensic investigators with clues to where the required potential evidence data can be found in the cloud.

The research question that this paper address is stated as follows: How can data provenance be used to provide digital forensic investigators with a more detailed layout of data in the cloud and identify the physical location in the cloud in an effort to help lessen the challenges the digital forensic investigation are faced with in cloud computing environments?

The remainder of this paper is structured as follows. Section 2 discusses the background of digital forensics, cloud computing and related work being done in cloud forensics. Section 3 discusses the requirements for a data provenance model that better suits the cloud. Section 4 is a discussion regarding the possible design of such a model, and section 5 discusses a possible practical implementation of such a system. Section 6 concludes with a short summary regarding the remaining work to be done.

II. Background

This section discuss an introduction of cloud computing and the challenges it holds for digital forensic investigators as well as related work carried out on this topic.

A. Cloud Computing

In this research paper we adopt the United States National Institute of Standards and Technology's (NIST) definition of cloud computing which defines cloud computing as: "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction."[3] The traditional model of computing viewed from a theoretical standpoint does not differ much from the model of cloud computing. The computing model in essence exists of input and output peripherals, one or more storage devices and a platform that connects these modules with one another. In a traditional computer these things can be classified as the keyboard and screen, hard-drives

978-1-4799-3383-9/14/$31.00 ©2014 IEEE

for storage and the motherboard and operating system serves as the platform to connect the devices. In the cloud computing environment it is much the same, the primary difference is in the fact that the storage device is no longer located in the same place as the input and output peripherals, and a network interface comes into play in order to access data on the storage devices. The model presented to end users of cloud services still look the same.

There are three categories of services, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [4]. IaaS involves hosting and development, PaaS is a platform configured for development and SaaS is fully functional software applications. If one breaks down the task a computer performs to: input, processing and output, then regardless of the category of cloud computing the end user is interested in, the processing is the same. The user provides some input, computations take place in the background and output is presented to the user. The actual location of data in the cloud is hidden from the users. This is a great advantage to normal cloud computing users but a tremendous challenge to digital forensic investigators.

B. Digital Forensics

Traditional computing saw the introduction of computer based criminal activities. Cloud computing are still faced with the same challenge of identifying and investigating computer based criminal activities. As mentioned in the introduction of this paper the traditional digital forensic investigation model does not scale well in the cloud. Corporate security teams do not have the freedom of performing independent investigations. The CSP has full control over the computing environment and thus also the sources of evidence [5]. Further more questions like who accessed specific data and information cannot be answered, if no corresponding logs are available. [5] writes that the history of a digital object, combined with a suitable authentication scheme is crucial information for a digital forensic investigation.

When referring to cloud security many researchers focus on isolating cloud instances [6][7] as well as the security of the hypervisor and network infrastructure [8], however some work has been done regarding digital provenance [9] and cloud forensics [10]. Many researchers have raised the need to have cryptographic proofs for verifying data integrity within the cloud as a requirement for digital forensics performed on cloud storage [11][12]. NIST proposed the design of Forensic Web Services that securely maintains transactional records between web services specifically related to web services that provide a monetary service to consumers [13].

From this work one can see a clear need for authenticating the integrity of data as well as the need to have data provenance available to provide the answers such as whom accessed a digital object, when and where it was accessed etc. Furthermore throughout the execution of a DFI it is of vital importance to maintain the chain of custody. In this work we will investigate how the chain of custody is kept in a DFI and how this can be applied within the infrastructure of the cloud to dynamically maintain a chain of custody on a digital object right from the start when the object enters the cloud. This however raises a few questions. The authors define the state of data in the cloud to be either at rest, in motion or in execution. At rest would signify data residing on a storage device, data in motion could either be transferring from storage to memory or could be transferring over a network from one device to another, data in execution is loaded in memory for the processor to handle and modify. When developing a service for maintaining data provenance one needs to consider what information is important, will a notation be made each time a digital object is transferred or accessed, or only if the object is changed?

Provenance information must be secure but must not violate the information confidentiality or privacy in cloud computing. There are a few requirements of data provenance as identified by [9], these include unforgeability which means no adversary should be able to forge a valid provenance record, and conditional privacy preservation which means that only a trusted authority may reveal the identity recorded in the provenance. Provenance information must be stored in a forensic ready manner meaning that the type of information that is captured and stored should be done in such a manner that the information is ready to be used in an investigation if one is required.

[9] proposed a model for secure data provenance in cloud computing environments that is based on the bilinear pairing technique [14] for providing trusted evidence for digital forensics. In this technique a group of users can be granted access to data in the cloud using anonymous authentication, which is a technique where a user is authenticated in a system without revealing the users identity. This is done through credentials computed for an entire group of users in such a manner that each user has unique credentials, but the entire groups credentials are related through mathematical inverses. Hence an entire group can be granted access to data in such a way that the exact user accessing the data is unknown. When a user connects to the system and the users credentials is presented, the system computes a private key for the user based on his credentials and the authorized group access key. If the user is authenticated as a member of the group, the system grants him access to the data in the cloud. This model provides provenance in a secure manner using this technique.

Another provenance system is proposed by [15], this system is designed specifically to provide digital provenance on objects by building components into provenance-aware network storage (NFS) that runs on top of a Provenance-Aware Storage System (PASS)

In the next subsection we look at the 7-layer OSI model. This model provides a good reference for network infrastructure.

C. 7-layer ISO OSI Reference model

The model, designed and developed from 1977 to July 1979 as architecture was designed due to the realization, at the time, that there existed a serious need for standards in computer

networks [16]. The OSI model exists of 7-layers, each one serving a specific purpose. If data is not required by a specific layer it is simply passed on to the next layer. Therefore each layer shields the other layers from complex data it does not need or have use for. The higher up in the layers one moves the more information becomes available for network analysis because the layer protocols transform the raw data into information that can easily be interpreted by investigators.

The application layer and the transport layer is of interest in this research. The Application layer is the highest layer in the OSI model and directly serves the end users purpose. Many user-defined protocols are associated with this layer and independent software applications run on this layer. The transport layer has the purpose of providing a pipe between two processes communicating over the network [17]. The transport layer also relieves the session layer from the detail of how reliable communication takes place between session entities. The protocols associated with this layer are the TCP and UDP protocols.

In the next section we look at establishing a clear set of requirements for a system to provide data provenance for objects in the cloud. We also investigate how such a system can be implemented to fit into the cloud computing architecture.

III. Requirements for data provenance in the cloud

From what has been identified in the literature and background section we now have a clear picture of what is expected from digital provenance. However we still need to identify where and when to capture provenance data.

The chain of custody as defined by [19] is only concerned with evidence from the time it was captured until it is presented in court. Therefore this does not help us determine what to capture for data provenance, however [20] states that the challenge that DFI examiners are faced with in the cloud is to determine the who, what, when, where, how, and why of cloud-based criminal activity. From this we can establish some requirements. Just as the chain of custody is concerned with keeping the integrity of data since it was collected, we need to keep the integrity of data in the cloud. Therefore it is necessary to capture provenance data if and when an object is modified. The what could be the object that was modified, when indicates the time of the modification, and finally where is the place in the cloud.

Theoretically this information should be relatively easy to capture. How an object was modified could be indicated by something like a hash code simply showing that something has changed, or it could indicate what application has modified the object. If the provenance record is extremely verbose the how may include the exact changes as one can see in a source-code version control log. Finally, who modified the object could most probably be the user logged into the system at that time, however it could also be a malicious program or a Trojan horse. Investigators are faced with an even bigger challenge in determining who accessed an object in the cloud, because of anonymous authentication. Anonymous authentication is an advantages feature for cloud users because it provides them with privacy. However the addition of this feature creates the need for provenance since it is no longer possible to establish with complete certainty, the user who accessed a specific object [9]. Therefore investigators need to be cautious in drawing any conclusions regarding who or what accessed a digital object. Investigators need to take into consideration anchor events when moving from digital to physical space [21].

In the rest of this section we list the requirements for a data provenance record as has been identified thus far. We then look at techniques used to present provenance data, followed by techniques for storing provenance data.

A. Requirements of a provenance record

From the requirements for identified by [9], the record has to have the following characteristics:
1. A provenance record need to be unforgeable.
2. A record need to be kept confidential.
3. The integrity of the record should be maintained by the system.
The record has to contain the following information:
4. Ability to answer the who, what, when, where, and how of an event.
4.1 Who: the identity of the process or user account associated with the modification.
4.2 What: the object that was modified.
4.3 When: the time of the occurrence.
4.4 Where: the object's location in the cloud at the time of the event.
4.5 How: The hash code of the object before and after the modification occurred.

B. Data provenance techniques

Data provenance records need to be captured and stored in a forensic ready manner. In this paper we are interested in the audit trail of a digital object. This trail can be provided by provenance data. There are two main approaches used to keep provenance data, known as annotations and inversion [22]. The inversion technique makes use of mathematical inverses that is defined from user-defined functions such as SQL queries. This technique are more compact than annotations but cannot be applied to all areas because not all user-defined functions has inverse functions. In this research we use annotations to represent provenance data. The annotation technique is much more flexible, and this provides the advantage that the format for representing and storing provenance data is also flexible. Many provenance systems using the annotation technique use XML to store the provenance data [23], [24]. Many service based architecture use XML as the primary format for message

exchange. The level of detail that is collected for provenance is scalable and should be based on the importance of the object.

C. Storing provenance data

Provenance data can be embedded with the data object and the object and its provenance stored as a single digital object, or the object and its provenance can be stored as two separate digital objects. Both methods have its pros and cons. Both the Flexible Image Transport System [25] and Spatial Data Transfer Standard [26] allows for metadata to be appended to the file's header. Appending the provenance to the header of the object helps to maintain the integrity of the provenance record because the record is kept with the object and can more easily be verified.

An alternative solution to store provenance data would be to separate the provenance data from the data object. The provenance data can still be stored in the same file-system as the data object or in a different system. [27] writes that the key to efficient digital forensic strategy is centralized logging. Logging data on a separate systems allows for easier access as well as better maintaining the integrity of the data because access to the data can be controlled through an effective access control mechanism.

IV. Capturing provenance records in cloud computing environments

Based on the requirements of this model and the network protocols available in the cloud, the question to be answered is, "Which protocols can best provide the information that will meet the requirements for provenance for a digital object?"Considering the purpose of OSI layers, if the investigator is concerned with how an object was routed, the network layer could best provide this information. However, capturing routing information will constitute considerable data that need to be stored because an annotation is created for each node the object passes. The question the investigator is concerned with in this research is the who, what, when, where, and how of an event. Taking this into consideration, the data provenance record should provide information regarding the creation of the object, and subsequent modifications.

Considering the integrity of an object as one of the core requirements of this model, it is necessary to keep a record of each modification to an object. In the OSI model, the only nodes that will have the ability to modify an object would be the source node where the object came from and the destination node where the object is being sent. Nodes in between only relay data objects to the destination.

Considering the state of an object defined by the authors as, at rest, in motion or in execution, and the purpose of this model, aimed at data provenance in the cloud. The goal of providing provenance is aimed at data in motion from one cloud node to another because investigators are often faced with the challenge of not knowing the physical location of data in the cloud. Therefore in this research the primary aim of provenance data is to provide the investigator with the information indicating the physical location of the data object in the cloud. The model will only maintain provenance for modification to an object while that object remains on a certain cloud node. It is necessary to determine who modified an object and when. Consider for an example an employees bank account details stored on file. If the file is modified and the account details changed with malicious intent, the employer may need to file criminal charges against the person responsible for the malicious activity.

A. Provenance data available from the OSI Layer Protocols

The goal of this section is to take a look at the protocols provided in the layers of the OSI reference model. These protocols provide information that can be used as part of the provenance data for digital objects. The authors are concerned with two of the OSI layers in this research. They are the Transport and Application layers. From the protocols available from these layers it is possible to get all the information required to meet the requirements for the model the authors propose.

4.2.1 Transport layer

From the TCP and UDP protocol headers it is possible to get the destination address where an object is being sent. This answers the where for that particular object. Although it should be noted that in this layer the address of a computer known as the IP-address is not necessarily fixed, but dynamic. The use of a dynamic address as an identifier has the advantage that the model scales better in the cloud because of the flexibility that this address provides. The digital forensic investigators can still rely on the logs provided by the cloud service providers to determine which physical machine was assigned a specific IP-address at a certain time.

4.2.2 Application layer

Considering that digital objects in the networking environments is transmitted in packets but investigators are concerned with application layer objects, the remaining questions namely who, what, when and how remains a challenge to be answered at the application layer. This requires the development and implementation of a digital forensic ready application or service to log and provide data provenance on application layer objects similar to the system proposed by [15]. The model will however differ from that of [15]. The architecture and applications of the [15] model is set up specifically in such a way to facilitate the model. The model proposed in this research needs to be flexible enough to apply to any cloud server without changing any of the architecture or applications and services that the server is currently providing. This compatibility can be provided with the addition of some software or service on the server.

V. Considerations for a practical implementation of a cloud based data provenance system

Suppose that metadata tags cannot be appended to the headers of all or any file type, the metadata tags should be stored separately from the data objects. Storing provenance

data on the same system with the data object would mean each time the object is transferred, the provenance record should be transferred with it. This presents a challenge. Storing the provenance data on a central server, not with the data object, would solve this challenge.

When an object in the cloud is modified, a provenance record containing details of the modification should be created. This requires application layer software on cloud servers to capture and create provenance records for maintaining data provenance in the cloud. [27] suggest the use of centralized logging to better maintain data integrity. Applying centralized logging to the cloud based provenance model would serve to maintain the integrity of the provenance records. Centralized logging would also solve the problem of having a digital object in the cloud that is transferred elsewhere and not knowing what to do with the provenance data. The digital object can simply be assigned an identification tag in the central provenance server and each time the object is modified somewhere in the cloud the record is sent to the server with all the necessary details and the identification tag to keep track of the object's audit trail through the cloud. There are some challenges in implementing this model in the cloud, irrespective of whether a central log server is used. These challenges include the development of an application layer protocol or standalone application that needs to be installed on cloud servers to create and transfer or maintain provenance records in the cloud. Privacy laws and policies also need to be considered when capturing data regarding details about the cloud users or the user?s data.

VI. Conclusion and remaining work

The goal of this research is to consider challenges digital forensic investigators are faced with in a cloud computing environment and to provide a solution that will ease a digital forensic investigation in the cloud through the application of a digital provenance system. In this research the authors propose a digital provenance system that captures data to answer the who, what, when, why and how of events in the cloud and to store this information on a central server to maintain the data integrity. The work that remains to be done includes the development and implementation of a prototype solution. The use of a central server for provenance records will facilitate easy access to provenance data for investigators. Consider one of the challenges digital forensic investigators face is the identification of the physical location of data. If provenance data can be stored on a centralized server the provenance record can indicate to the investigator the physical location of the data object at a specific point in time.

VII. Acknowledgement

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Numbers 88211, 89143 and TP13081227420)

References

[1] G. Yong Ph.D, "Digital forensics: Research challenges and open problems," in *Department of Electrical and Computer Engineering and Information Assurance Center Iowa State University*, December 2007.

[2] D. Birk and C. Wegener, "Technical issues of forensic investigations in cloud computing environments," in *Ruhr-University Bochum, Horst Goertz Institute for IT Security Bochum, Germany*, January 2011.

[3] P. Mell and T. Grance, "The nist definition of cloud computing," in *Information Technology Laboratory - Computer Security Division*, September 2011.

[4] Y. Jadeja and K. Modi, "Cloud computing - concepts, architecture and challenges," *Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on*, pp. 877 –880, March 2012.

[5] D. Birk and C. Wegener, "Technical issues of forensic investigations in cloud computing environments," in *Systematic Approaches to Digital Forensic Engineering (SADFE), 2011 IEEE Sixth International Workshop on*. IEEE, 2011, pp. 1–10.

[6] W. Delport, M. Kohn, and M. S. Olivier, "Isolating a cloud instance for a digital forensic investigation," in *Proceedings of the 2011 Information Security South Africa (ISSA 2011) Conference*, Johannesburg, South-Africa, August 2011.

[7] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds," in *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 199–212.

[8] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling data in the cloud: outsourcing computation without outsourcing control," in *Proceedings of the 2009 ACM workshop on Cloud computing security*. ACM, 2009, pp. 85–90.

[9] R. Lu, X. Lin, X. Liang, and X. S. Shen, "Secure provenance: the essential of bread and butter of data forensics in cloud computing," in *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*. ACM, 2010, pp. 282–292.

[10] K.-K. Muniswamy-Reddy and M. Seltzer, "Provenance as first class cloud data," *ACM SIGOPS Operating Systems Review*, vol. 43, no. 4, pp. 11–16, 2010.

[11] Y. Shi, K. Zhang, and Q. Li, "A new data integrity verification mechanism for saas," in *Web Information Systems and Mining*. Springer, 2010, pp. 236–243.

[12] A. Juels and B. S. Kaliski Jr, "Pors: Proofs of retrievability for large files," in *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 2007, pp. 584–597.

[13] M. Gunestas, D. Wijesekera, and A. Singhal, "Forensic web services," in *Advances in Digital Forensics IV*. Springer, 2008, pp. 163–176.

[14] X. Boyen and B. Waters, "Full-domain subgroup hiding and constant-size group signatures," in *Public Key Cryptography–PKC 2007*. Springer, 2007, pp. 1–15.

[15] K.-K. Muniswamy-Reddy, U. Braun, D. A. Holland, P. Macko, D. Maclean, D. Margo, M. Seltzer, and R. Smogor, "Layering in provenance systems," in *Proceedings of the 2009 USENIX Annual Technical Conference*, 2009.

[16] H. Zimmermann, "Osi reference model–the iso model of architecture for open systems interconnection," *Communications, IEEE Transactions on*, vol. 28, no. 4, pp. 425–432, 1980.

[17] M. Olivier, *Computer Network Fundamentals a Local Topdown Approach*, University of Pretoria, 2011.

[18] G. A. Deaton Jr and R. O. Hippert Jr, "X. 25 and related recommendations in ibm products," *IBM systems journal*, vol. 22, no. 1.2, pp. 11–29, 1983.

[19] W. Zinnikas, "Chain-of-custody considerations," 2003.

[20] J. J. Barbara, "Cloud computing: Another digital forensic challenge," *Digital Forensic Investigator News*, October 2009.

[21] P. M. Trenwith and H. Venter, "Digital forensic readiness in the cloud," in *Information Security for South Africa, 2013*. IEEE, 2013, pp. 1–5.

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

[23] J. Zhao, C. Wroe, C. Goble, R. Stevens, D. Quan, and M. Greenwood, "Using semantic web technologies for representing e-science provenance," in *The Semantic Web–ISWC 2004*. Springer, 2004, pp. 92–106.

[24] R. Bose and J. Frew, "Composing lineage metadata with xml for custom satellite-derived data products," in *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*. IEEE, 2004, pp. 275–284.
[25] R. Hanisch, A. Farris, E. Greisen, W. Pence, B. Schlesinger, P. Teuben, R. Thompson, and A. Warnock, "Definition of the flexible image transport system (fits)," *Astronomy and Astrophysics*, 2000.
[26] P. Altheide, "Spatial data transfer standard (sdts)," in *Encyclopedia of GIS*. Springer, 2008, pp. 1087–1095.
[27] J. Tan, "Forensic readiness," *Cambridge, MA:@ Stake*, 2001.