International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 3 Issue 4 April 2016

Data Provenance - An Essential Ingredient in Cloud Forensics Investigation

Pooja A R[1], Raj Kumar T[2]
Department of Computer Science and Engineering, College of Engineering Kallooppara, Kerala, India[1, 2]
pooja.a.r.1992@gmail.com[1], rajcek@gmail.com[2]

Abstract: *The cloud is going to become the next computing environment in both data storage and computation, due to its pay-as-you-go and provision-as-you-go models. Cloud storage plays great role in several cases such as to back up desktop user data, used to host shared scientific data, used to store web application data and also used to serve web pages. But in today's cloud stores there is an essential ingredient is missing that is data provenance. Provenance is the metadata of an object that describes the history of that particular object. With the use of provenance, data users can check the identity or authenticity of data of interest. Data provenance will play a significant role in cloud forensics investigation in future.*

Keywords: *Provenance, Cloud computing, Virtualisation, Cloud forensics.*

1. INTRODUCTION

Cloud computing [1] is one of the backbone of many services. Now a days, users are enjoying the services provided by clouds when they access Gmail, Google Calendar, Dropbox, SkyDrive, Social medias or run hundreds of Amazon Elastic Compute Cloud (EC2) instances for processing large-scale data. In spite of that, these computing and storage resources provided by the cloud can be abused by intruders to perform attacks from machines inside the cloud. A malicious user can also store some secret files in cloud storage to keep their PC clean. To investigate these types of crimes in clouds, investigators have to perform a digital forensic investigation in the cloud environment.

Miserably, many of the assumptions of conventional digital forensics are not at all valid in the cloud computing model. One of the major problem is that neither users or nor investigators have direct access to the cloud. In case of cloud environment, each cloud server contains files from many users. Hence, it is not viable to grab servers from a data center without breaching the privacy of many other users. The credibility of the evidence is also doubtful, because other than the cloud service provider's word, there is no usual way to decide the veracity of the evidence. To furnish on-demand services, cloud providers do not encourage persistent storage for terminated VMs. Hence, data lies in the cloud VMs will be inaccessible after terminating the VMs. This in turns makes it almost impossible to do forensics investigation if some illegal activities have been occurred using such terminated VMs. To solve all these issues, cloud provenance will play a great role. Provenance improves the value of the data and it describes how was an object is created, difference in ancestries of two objects and on what other objects does an object depends etc. Data provenance is a type of metadata that pertains to derivation history of a data starting from its source of origin.

2. BACKGROUND AND LITERATURE REVIEW

2.1 Cloud Computing

Cloud computing is the distribution of computing services over the Internet. Cloud services permits individual users and businesses to use software and hardware that are controlled by third parties at distant locations. Examples of cloud services

contains online file storage, Social Medias, webmail, and online business applications. The cloud computing model provides access to information and computer resources from anyplace that a network connection is accessible. Cloud computing offers a shared pool of resources, including data storage space, networks, computer processing power, and dedicated corporate and user applications.

The following description of cloud computing has been refined by the U.S. National Institute of Standards and Technology (NIST) [2].

"*Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models*."

2.2 Characteristics

The characteristics of cloud computing [3] contains on demand self-service, broad network access, resource pooling, rapid elasticity and measured service. On demand self-service means that consumers can manage their own computing resources by requesting the cloud data. Broad network access permits services to be obtainable over the Internet or private networks. Pooled resources means that consumers draw from a pool of computing resources, commonly in remote data-centers. Services can be scaled higher or lesser; and use of a service is measured and consumers are billed accordingly.

2.3 Service Models

The services provided by the cloud service providers can be categorized into three models.

Software as a Service (SaaS): In this model, a whole application is offered to the customer, as a service on demand. A single instance of the service runs on the cloud and multiple end users are serviced on the customer's side. Nowadays SaaS is offered by enterprises such as Google, Microsoft, Oracle's on demand, Abiquo, Zoho, etc.

Platform as a Service (Paas): In this model a layer of software, or development environment is encapsulated and offered as a service, upon which other higher levels of service can be made. To meet manageability and scalability requirements of the applications, PaaS providers offer a predefined mixture of OS and application servers, such as LAMP platform (Linux, Apache, MySQL and PHP), restricted J2EE, Ruby etc. Google'sApp Engine, Force.com, etc are some of the popular PaaS examples.

Infrastructure as a Service (Iaas): IaaS mainly offers basic storage and computing capabilities as standardized services over the network. Servers, storage systems, networking infrastructure, data center area etc are shared and made accessible to manage workloads. The user would normally install his own software on the infrastructure. Some common examples are Amazon, GoGrid, at&t, rackspace 3 Tera, etc. Figure1 shows the three service models of cloud computing.

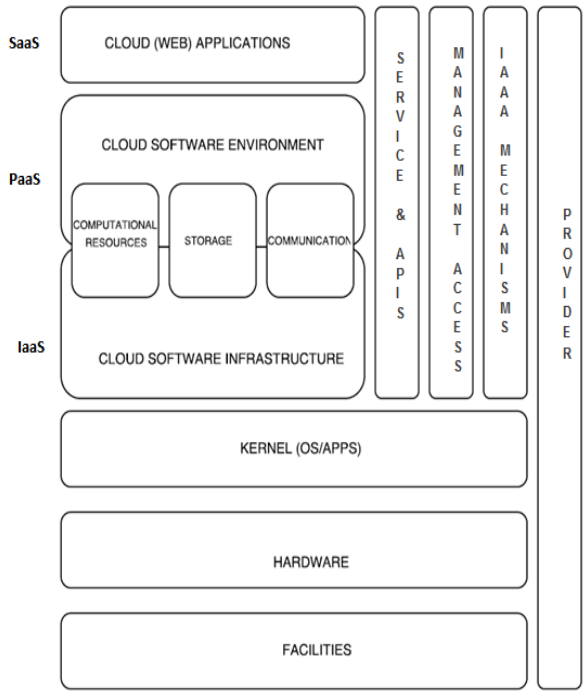

Fig.1 Three service models of Cloud Computing

Figure 2 shows the limited amount of customers control in different layers for the three service models – IaaS, PaaS, and SaaS. In IaaS, customers have additional control over resources than SaaS or PaaS. The less control over resources has made the data collection in SaaS and PaaS more challenging than in IaaS. It is even difficult in some cases. If

it's possible to get the image of an IaaS instance, it will be easy to investigate the system. For SaaS and PaaS, need to depend on the Cloud Service Provider. User can only get a high level of logging information from this two service models. As users have control over the application installed in PaaS, they can preserve log of various actions to enable the investigation procedure. In case of SaaS, customers really have no control to log the actions.

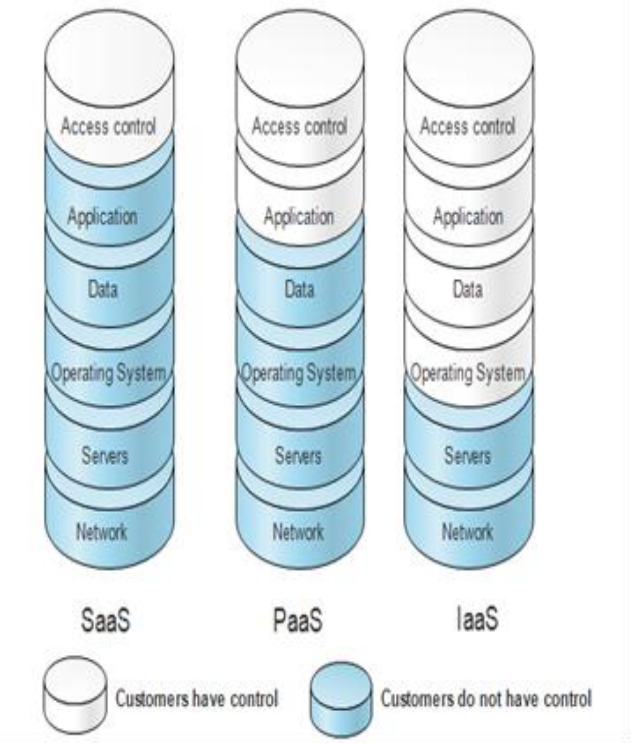

Fig.2 Users control over SaaS, PaaS and IaaS

2.4 Deployment model

Based on the deployment model cloud computing can be classified into three:

Public Cloud : Public clouds are maintained and operated by third parties; they deliver superior economy of scale to customers, as the infrastructure costs are spread among a mix of users, giving each individual client an attractive low-cost, Pay-as-you-go model. Many customers can share the same infrastructure pool with limited configuration, security protections, and accessibility modifications. These all are attained and supported by the cloud service provider. One of the advantages of a Public cloud may be larger than an enterprises cloud, thus providing the ability to scale seamlessly, on demand.

Some of the public cloud providers are listed below.

- AT&T
- Amazon AWS
- Rackspace
- IBM
- HP
- Microsoft cloud services
- Nirvanix

Private Cloud: Private clouds are constructed entirely for a particular enterprise. Their intention is to address concerns on data security and provide greater control, which is typically lacking in a public cloud. A private cloud offers the similar basic benefits of public cloud. It include self-service and scalability, multi-tenancy, the capacity to provision machines, altering computing resources on-demand, and producing multiple machines for complex computing jobs, such as big data.

Private cloud computing is a style of computing in which scalable and elastic IT enabled capabilities are delivered as a service to customers by the organizations. Cloud-computing services can be delivered by an internal IT organization or by an external service provider. The fundamental infrastructure can be hosted within an organization's data center or in an external data center. That infrastructure can be dedicated to a single customer, shared between a consortiums of customers or shared with a service provider's customer base in general. Private cloud computing can be in many forms [5]:

- A private cloud will usually be insourced and run on-premises using equipment owned by the enterprise.
- Private cloud services will usually be IaaS, but not always.
- A private cloud that is IaaS will usually influence virtual machines.

The main benefit of private cloud computing is speed. Implementing a service catalogue that offers standard services through a self-service interface, and automating the delivery of those offerings, can increase the speed of delivery

dramatically. By themselves, standards, automation, and some form of resource pooling or virtualization will also reduce costs. Many vendors provide cloud management software to build public and private IaaS cloud.

Some of the private cloud providers are listed below:

- VMware vcloud suite
- Amazon AWS
- Open stack(open source)
- Cloudstack
- Rackspace
- Microsoft cloud services

Hybrid Clouds: It combines both public and private cloud models. With a Hybrid Cloud, service providers can consume third party Cloud Providers in a full or partial manner. It increases the elasticity of computing. The Hybrid cloud environment is able to provide on-demand, externally provisioned scale. The ability to augment a private cloud with the resources of a public cloud can be used to manage any unexpected rushes in workload.

Some of the hybrid cloud providers are listed below.

- VMware
- Amazon AWS
- EMC
- HP
- Rackspace
- Microsoft cloud services

3. PROVENANCE IN CLOUD

Provenance [6] is the metadata that describes the ancestry of an object, it is also known as lineage. Without provenance, data clients have no resources to validate genuineness or identity of a cloud data. Digital provenance is metadata that defines the source or history of a digital object. Provenance improves the significance of the data it describes. And also it provides answers to various questions such as.

How an object was is created?
On what other objects does this particular object depend?
How do the histories of these two objects differ?

3.1 Uses of Provenance

- Provenance [7] can provide hints on access patterns, detect anomalous behaviour, and provide enhanced user search capabilities
- Provenance is crucial in a number of areas such as scientific computation, security, regulatory compliance, and data archival.
- Scientists use provenance for experimental reproducibility or to decide what changes are occurred between two runs of same experiment.
- In case of security, it can be used to validate how a virus spread through a system.
- In business domain provenance is used to prove evidences about information leak.
- Archivists maintain provenance meta-data to support document viability, renderability, understandability, authenticity, and identity in preservation contexts.
- Cloud stores can use provenance to build models of both applications and users

4. EXISTING SYSTEM

The Provenance Aware Service Oriented Architecture (PASOA) [8] project is used to build a provenance infrastructure for recording, storing and reasoning over provenance using an open provenance protocol that will foster interoperability among e-science communities. This project identifies several requirements for a provenance system in a service oriented architecture, such as verifiability of actors involved in a process and reproducibility of the process, accountability and preservation of provenance over time, scalability of the provenance system, generality to support diverse Grid applications as well as customizability as required. Actors are either clients of services or services that are invoked, and they generate two kinds of provenance during a workflow's execution. Interaction provenance, describing the input and output parameters of a service invocation, is produced and verified by both actors – client and service – in the invocation. Actor provenance includes

metadata about the actor's own state during a service invocation and is not provable.

The Provenance Recording Protocol (PReP) defines fourteen interaction provenance messages that are generated by the actors, synchronously or asynchronously, with each service invocation. They are divided into four phases: negotiation phase, invocation phase, provenance recording phase, and termination phase, during which the actors agree upon a provenance service to record the provenance, perform the service invocation, record their interaction provenance, and terminate the protocol, respectively. All interaction and actor provenance messages produced by the clients and services in a workflow are linked using an ActivityID exists in the provenance messages. These form a process oriented provenance trace of the workflow recorded as annotations, and data provenance needs to be individually derived by joining all statements that have the similar ActivityID as that of the statements having the data as output.

Provenance Recording for Services (PReServ) is a web service implementation of the PReP protocol that stores the provenance either in memory, in a relational database, or in the file system. The actual representation of provenance is not apparent. A performance overhead of 10% has been observed when the provenance assertions are given in to asynchronously by the actors, with the overhead increasing when the provenance is sent synchronously with the service invocation. Overhead also lies in altering the actors to produce the provenance messages, though this can be reduced by just adjusting the workflow engine to produce provenance. Methods to scale the provenance store through association are being considered.

A querying interface is not defined as part of the PReP protocol but a basic querying API is available to retrieve provenance from PReServ. Basic queries to trace all data that were derived using the same service can be accomplished. Semantic validity checking of services and their inputs/outputs is possible by comparing the expected inputs/outputs of a service, available in a semantic registry, with the actual inputs/outputs available with the provenance. Many other uses are there, such as repeating a workflow using the inputs to services available as provenance, are also predicted.

5. CONCLUSION

As the use of cloud computing is increasing day by day, there is an increase in the significance of providing trustworthy cloud forensics methods. Provenance enhances the value of the data and it enables cloud consumers, law enforcement officers, and forensic investigators to acquire trustworthy forensic data provenance independent of the cloud provider. If a malicious user stores some secret files in cloud storage to keep their PC clean after performing a cybercrime using that secret files. To investigate these types of crimes in clouds, investigators have to perform a digital forensic investigation in the cloud environment. Miserably, many of the assumptions of conventional digital forensics are not at all valid in the cloud computing model. By using provenance data, an investigator can able to filter out unwanted data by checking the history of data in the cloud using the given evidences. So data provenance will play a great role in forensic investigation in cloud forensics in future.

ACKNOWLEDGEMENTS

Many people have contributed to the success of this. Although a single sentence hardly suffices, I would like to thank Almighty God for blessing us with His grace. I extend my sincere and heartfelt thanks to Mr.Harikrisha S for providing us the right information at right time to carry out this work. Last but not the least, I thank all others, and especially my classmates who in one way or another helped me in the successful completion of this work.

REFERENCES

[1]. S Zawoad, R Hasan. FECloud: A Trustworthy Forensics-Enabled Cloud Architecture 2015, 11th Annual IFIP WG 11.9 International Conference on Digital Forensics, Orlando, Florida, January 2015.
[2]. NIST, (2012), "Definition of Cloud Computing," [Online].Available:http://csrc.nist.gov/groups/SNS/cloud computing/index.html. [Accessed 09 2012].
[3]. TorryHarris,"Cloud computing-An Overview, Whitepaper 2010.
[4]. Saibharath S and G Geethakumari, "Design and Implementation of a forensic framework for Cloud in Openstack cloud platform.Proceedings of the IEEE International Symposium on Cloud Computing: Architecture,

Applications and Approaches (CCA-2014), September 24-27, India, 2014.
[5]. Thomas J. Bittman. ―Private Cloud Computing: An Essential Overview, Gartner 2013.
[6]. Muniswamy-Reddy, K.-K., Macko, P., And Seltzer, M. Provenance for the cloud. In FAST '10: Proceedings of the 8th USENIX conference on File and storage technologies (2010), USENIX, pp. 15–14.
[7]. Olive Qing Zhang, Markus Kirchberg, Ryan K L Ko, and Bu Sung Lee: How to Track Your Data: The Case for Cloud Computing Provenance.
[8]. Yogesh L. Simmhan, Beth Plale, Dennis Gannon, "ASurvey of Data Provenance Techniques".