From System-centric to Data-centric Logging – Accountability, Trust & Security in Cloud Computing

Ryan K. L. Ko, Markus Kirchberg, Bu Sung Lee
Service Platform Lab, Cloud & Security Lab
Hewlett-Packard Laboratories Singapore
Fusionopolis, Singapore
{ryan.ko | markus.kirchberg | francis.lee}@hp.com

Abstract—**Cloud computing signifies a paradigm shift from owning computing systems to buying computing services. As a result of this paradigm shift, many key concerns such as the transparency of data transfer and access within the cloud, and the lack of clarity in data ownership were surfaced. To address these concerns, we propose a new way of approaching traditional security and trust problems: To adopt a detective, data-centric thinking instead of the classical preventive, system-centric thinking. While classical preventive approaches are useful, they play a catch-up game; often do not address the problems (i.e. data accountability, data retention, etc) directly. In this paper, we propose a data-centric, detective approach to increase trust and security of data in the cloud. Our framework, known as TrustCloud, contains a suite of techniques that address cloud security, trust and accountability from a detective approach at all levels of granularity. TrustCloud also extends detective techniques to policies and regulations governing IT systems.**

Keywords—trust; cloud computing; cloud computing security; data-centric logging; accountability; TrustCloud framework.

I. Introduction

The recent increase in uptake of cloud computing signifies an underlying paradigm shift in computing technology – a shift from *systems-as-assets* to *information-as-assets* [1].

A. A Paradigm Shift in Technology

The emergence and rapid adoption of cloud computing has brought about a paradigm shift in how individuals, businesses and organizations (in short, users) utilize and think about computing resources. The whole IT infrastructure has become available as a service; thus, mitigating the need to own, maintain or deal with the risk of hardware and security of IT infrastructure. This shift enables a much more efficient use of computing resources, but places the emphasis on the need for more adequate tools and mechanisms to safeguard the integrity and accuracy of data and information, the real assets in the cloud era. With cloud computing, enterprises are empowered to focus on value-added activities such as generating sales and providing services and no longer need to allocate resources to procuring and managing systems. All cloud computing users need to do is to decide how much computing and storage they require and pay directly for the services [2]. As such, users no longer 'own' systems, but only their data. While convenience is increased, the reduction of ownership over the machines, which store the data, decreases the overall sense of control and 'trust', especially when it involves sensitive data such as healthcare, government, or financial data. This lack of ownership and transparency are some of the main obstacles for data-sensitive industries to 'trust' the cloud [3]. The question now is whether such loss of control can be mitigated and aided by an increased awareness of *who* has touched the data, *how* it was touched, at *where* and *when*.

While many providers claim that the data is still owned by the end-users, the definition of data ownership is a contentious subject by itself [4]. When we place the data into a provider's cloud, how can we ensure that the provider cannot and will not access or mine the data for their own use? For example, Google mines the contents of their Gmail end-users' emails in order to recommend each end-user specific context-aware advertisements into their advertising space alongside the mailboxes. While it provides Gmail for free, Google makes some of its revenues from advertising that depends on the content of the user's messages [5]. Does that mean that Google need to own or co-own the data first before it can process them for its own revenue? Also, are current laws and regulations enough to protect privacy, or deter malicious insiders from abusing the data housed within clouds? These questions lead us to consider the relevancy and currency of existing regulations and security tools. The best way to evaluate this is to reflect on the recent cloud security incidents.

B. Relevancy of Current Security Approaches

1) Recent High-profile Security Incidents

a) Google's Lost Email Accounts

In February 2011, 150,000 Gmail accounts were 'accidentally reset' by Google, leaving the users with all their previous messages wiped out from their mail [6]. This incident led many reports to highlight the importance of backing up data, even if it is stored in the cloud.

b) EMC/ RSA Security Breach

In March 2011, RSA, the security component of EMC suffered a data breach where attackers apparently stole intellectual property relating to the company's two-factor authentication technology [7]. A few months after that incident, RSA admitted that defense contractor Lockheed Martin was attacked using the stolen information from March 2011. At the time of writing (June 2011), RSA has offered to replace SecurID tokens for a certain subset of its customers, but it still has not admitted what information was stolen. This led to forums and blogs voicing out customers' concern about how protected they are, as the SecurID token is widely used by banks, large organizations and government agencies.

c) User Data Stolen from Sony Playstation Network

In April 2011, Sony admitted that its Playstation Network has been intruded, leaving the user names, passwords, addresses, birth dates, and other information used to register accounts compromised [8]. The stolen information may also include payment-card data, purchase history, billing addresses, and security answers used to change passwords, leading to the possibility of future identify theft scams.

d) UK National Healthcare System (NHS) Hacked

In June 2011, the IT systems of the UK's National Healthcare System (NHS) were hacked [9]. While the hackers uncovered several admin account passwords, health officials played down the security lapse and insisted that it affected only local systems and that no patient records were accessed. These incidents sparked fear of whether patient data stored in the NHS systems have been compromised.

e) Amazon EC2 Outages

In April 2011, Amazon, one of the top cloud computing providers, suffered an outage to its EC2 services [10]. The outage was caused by a configuration error during a network upgrade; sending a number of prominent cloud-based Web-sites offline, including Quora, Foursquare and Reddit. The above incidents resulted in a growing call for increased accountability of cloud service providers.

2) Preventive versus Detective Approaches

Trust components can be classified as ***Preventive Controls*** or ***Detective Controls***. Preventive controls are used to mitigate the occurrence of an action from continuing or taking place at all (e.g. better encryption techniques, or network and host firewalls blocking all but allowable activity). Detective controls are used to identify the occurrence of a privacy or security risk that violates policies and procedures (e.g. an intrusion detection system, or security audit trails, logs and analysis tools). Detective approaches also act as psychological obstacles to go against policies and/or procedures in the cloud, and also serve as a record for post-mortem forensic investigations should any non-compliance occur.

If we look at current cloud security approaches (e.g. end-to-end data encryption), we can observe that they mainly focus on preventive approaches. The high-profile security breaches above have demonstrated that current preventive methods are insufficient, as they do not provide users the transparency and accountability of their services. The preventive approaches also do not help the users achieve audit trails of data-centric events. As shown in the security breaches discussed, end-users who entrust their sensitive data in the providers' environments are often left disappointed, and worse, unable to attain full accountability concerning their data or data about them.

Hence, in this paper, we call for the increased deployment of detective approaches to improve trust and security in cloud computing. Preventive approaches are necessary but not sufficient. On top of increased deployment, we need to also improve current detective approaches. An analogy for the rationale of improving detective approaches can be drawn from the advancement of criminal forensics from thumbprint identification to the identification of DNA.

3) Data-centric versus System-centric Logging

As mentioned in Section I.A, traditional in-house controls (i.e. physical, logical and personal) are no longer valid when moving data and information assets into the cloud. New means of ***data-centric*** governance and compliance, accountability, protection from loss, corruption, and destruction etc. are required, in particular in the context of virtualization, which enables the cloud elasticity promise. Virtualization, however, means that accountability might require the identification not only of the virtual server in which events take place, but also the physical server. Currently, there is still a lack of transparency of (1) the linkages between the virtual and physical systems, (2) relationships between virtual locations and physical static server locations, and (3) how the files are written into both virtual and physical memory addresses. Such information is currently not available as a single-point-of-view. From a system's perspective, data exists at the level of blocks and files and directories. While there might be a multitude of operating systems (OSs) deployed in a single cloud, the majority of such OSs have not been designed for the cloud. In particular, traditional logging is process and/or event-based (for a particular user or node). In the cloud, however, there are no clear user or node barriers; instead, logging should be done with respect to the key assets, i.e., data and information. In terms of OSs, this means ***data-centric logging***. By data-centric, we refer to the need to trace data and files from the time they are created to the time they are destroyed; thus, data and information is viewed independent from the environmental constraints. This is reflective of the very elastic nature of cloud computing. With the transfer of control of data into the cloud, the providers have the mandate to ease the minds of consumers by facilitating them with the capabilities to track their assets.

II. Requirements for a Data-centric Logging for Increased Cloud Accountability and Trust

We envision data-centric logging to be able to achieve the following requirements:

A. Tracking Files

Within data-centric detective approaches, file-centric logging enables us to know, trace and record the exact file life cycle. This requires tracking of system read/write calls to the underlying file system. From the system read/write calls, we can also extract the files' virtual and physical memory locations, providing more information for further forensics. In the cloud, the file-centric perspective, however, must not be limited to a single node; as clouds are vast networks of physical and virtual servers over a large number of locations, we need to also monitor network logs within the cloud in order to be able to capture movement of a file from one node to another (be it a virtual or physical node) as well as the exit/entry point(s) of a file into/from the cloud.

Besides knowing about the travelling history of a file as well as access and modification details, file-centric logging is vital when enabling the "replay" of a snapshot, i.e. a reproduction of the exact state of the cloud at a particular moment. With a large number of virtual machines turned on and off at different time periods, and the execution of various business applications at the same time, it is very difficult to replay the exact same snapshot of the Cloud from the past, e.g. 1 hour ago, so that one can determine what actually went wrong [1]. File-centric tracking mechanisms log the resources the VMs use and share when they are turned on. Evidently, such snapshots cannot be captured in the data, information and workflow layers as they are too high-level and dependent on the on and off status of their hosting machines.

B. Tracking Data

The file abstraction enables us to tackle the snapshot replay issue in the cloud, but it is not a suitable abstraction from an informational point of view. In order to enable reasoning about the origins, collection or creation, evolution, and use of data, it is essential to track the history of data, i.e., its provenance. Provenance information is commonly seen as the foundation for any reasonable model of privacy and trust as it enables validation of the processes involved in generating/obtaining the data and the detection of unusual behavior. While these advantages are very promising, corresponding challenges are equally difficult to address. Such challenges include efficiently and effectively managing the sheer amount of provenance data that has to be maintained; ensuring consistency and completeness of provenance data; detecting malicious users who attempt to falsify provenance data; protecting data owner as well as data providers from exposing sensitive, confidential, proprietary or competitively important information indirectly through provenance logs; enabling efficient querying of provenance data; etc.

Besides provenance, other key concerns mandating data-centric logging include the need for support of consistency assurance, rollback, recovery, replay, backup, and restoring of data. Such functionality is usually enabled by using operational and/or transactional logs. Such logs have also been proven useful for monitoring of operational anomalies. While these concepts are well established in the database domain, cloud computing's characteristics such as eventual consistency, 'unlimited' scale, and multi-tenancy pose new challenges. In addition, secure and privacy-aware mechanisms must be devised not only for consistency logs but also for their backups, which are commonly used for media/node recovery.

C. Tracking Information

While data represents raw facts, it is information, which is derived from data that is of most interest and value to individuals, businesses and organizations. Information is used to reveal the meaning from data; obtaining timely and relevant information is key to good decision making and sustained success of most businesses. Key challenges to be addressed by logs are how information is derived from data and how information evolves in coherence with respect to the underlying data.

D. Tracking Information and Data Flows

Sections II.B and II.C are mainly concerned with individual aspects of data and information. We have yet to cover the workflow and business rule context in which data and information are typically embedded. That is, the need for audit trails and the audit-related data found in the software services in the cloud. High level fraudulent risks such as procurement approval routes, decision making flows and role management in software services have to be monitored and controlled. Hence, accountability of services/business functions and their providers within the cloud have to be managed. However, achieving auditability via methods such as continuous auditing [11] within a highly virtualized environment is a very difficult and complex task. There needs to be considerations for not only the auditing of the business logic and control flows, but also the applications implementing them.

III. The TrustCloud Framework – Data-centric Abstraction Layers

One of the most common groupings or layers in cloud computing is the view of Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). However, given the requirements listed in Section II, we now observe that these abstractions are mainly system-centric. In contrast, the TrustCloud framework takes a different perspective, i.e., an architectural, data-centric view. This results in a separate set of layers, which can accurately encompass and describe the scope of data-centric logs. Because of the scale of cloud computing, the types of data-centric logs range from *system-level* file-centric logs to *workflow-level* audit trail logs.

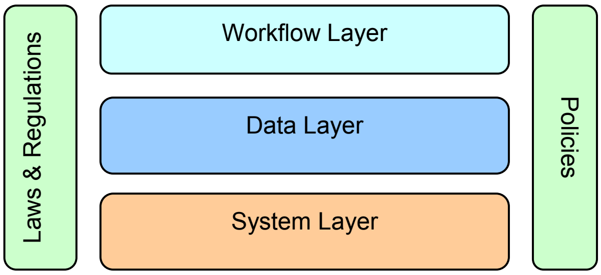

Figure 1. The TrustCloud framework [1]

There needs to be a clear definition of abstraction layers to reduce ambiguity and increase research focus and impact. The TrustCloud framework, proposed in our earlier works [1, 3], attempts to describe the layers of cloud accountability shown in Figure 1. This figure shows the five abstraction layers for the types of logs needed for an accountable cloud:

1. *System layer* – addresses tracking of files across the cloud.
2. *Data layer* – addresses tracking of change of data and information across the cloud.
3. *Workflow layer* – addresses data and information flow in the cloud.
4. *Law and regulations layer* – addresses data-centric logging requirements mandated by external laws and regulations.
5. *Policies layer* – addresses data-centric audit requirements mandated by internal governance and audit requirements.

It is important to note that the focus is on the abstraction layers of logs and not on architectural layers. Hence, the TrustCloud framework is independent of virtual or physical environments. Such explicit definition of layers allows us to efficiently identify the areas of their application and their focus areas. Further details of the TrustCloud Framework can be found in [1, 3].

IV. Related Detective Approaches

This section surveys current detective approaches.

A. TripWire

TripWire [12] creates a baseline database of all hash keys of files, then checks them at a schedule stated by a user, when hash keys are changed, the files affected are highlighted. Hence, it mainly identifies files that are changed via a comparison approach, i.e. before versus after. When some intentional user changes are made to the files tracked by TripWire, the user has to consciously update the baseline database. While TripWire is a popular intrusion detection tool, it is a user space application; hence, its scope is not able to cover the kernel level vulnerabilities and intrusions. This is a concern when we consider the cloud, which has kernel spaces in both virtual and physical machines. At the same time, the need for updating the key database regularly also means that such a technique is not scalable for the dynamic environments in cloud computing. Also, as the detection is on whether a change occurred, rather than the actual history of change, there is also no provenance recorded. This means that this is a best-effort data-centric detective method.

B. HyTrust Appliance

Recently in 2010, HyTrust [13], a startup focusing on cloud auditing and accountability, has released a hypervisor consolidated log report and policy enforcement tool *(i.e. HyTrust Appliance)* for virtual machine accountability management in clouds. In the context of Section III, HyTrust Appliance addresses the *System layer* of accountability in the cloud. Despite this, it focuses on the virtual machine layers and did not mention capabilities for virtual-to-physical complexities. Also, it views logging for accountability from system-centric perspective and not a file-centric perspective.

V. Future Work

Our team is actively researching on the following data-centric research challenges:

- Data provenance awareness across both virtual machines and physical machines in the cloud.
- Efficient methods for storage of cloud data-centric logs.
- Visualisation of cloud data-centric logs.
- Summarising and reporting anomalies mined from cloud data-centric logs.

VI. Conclusion

The objective of this paper is to encourage the adoption of file-centric and data-centric logging mechanisms as a means to increasing accountability, trust and security in cloud computing. The urgency for data-centric detective approaches was also discussed over a survey of recent high-profile cloud-related security breaches. Data-centric logging not only allows transparency of the movements of data in the cloud, but also addresses the growing concern of accountability of cloud service providers and provides pointed information for data leakage protection (DLP) and information life cycle management. As a detective approach, data-centric logging complements the prevalent preventive security approaches such as end-to-end data encryption. The approach provides records of data provenance and meaningful trails revolving around the life cycle and transfer of data, the most valuable assets of cloud end-users. We reiterated the importance of viewing cloud accountability via the five proposed levels: *system, data, workflow, laws and regulations, policies.* Related detective approaches were also reviewed but found lacking provenance features or comprehensive coverage of both virtual and physical servers typically found in cloud environments. Our team is currently actively researching on the key areas around the problem of data-centric detective approaches for cloud accountability.

Acknowledgment

The authors wish to acknowledge their TrustCloud team members: Peter Jagadpramana, Suen Chun Hui, Alan Tan, Lim Teck Hooi, Miranda Mowbray, Siani Pearson, Ahmed Aneeth, and Ahmed Rifau Rasheed.

References

[1] R.K.L. Ko, P. Jagadpramana, M. Mowbray, S. Pearson, M. Kirchberg, Q. Liang and B.S. Lee, "TrustCloud - A Framework for Accountability and Trust in Cloud Computing," *Proc. IEEE 2nd Cloud Forum for Practitioners (IEEE ICFP 2011)*, IEEE Computer Society, 2011, pp. 1-5.

[2] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin and I. Stoica, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, 2010, pp. 50-58.

[3] R.K.L. Ko, B.S. Lee and S. Pearson, "Towards Achieving Accountability, Auditability and Trust in Cloud Computing," *Proc. International workshop on Cloud Computing: Architecture, Algorithms and Applications (CloudComp2011)*, Springer, 2011, pp. 5-18.

[4] C. Reed, "Information 'Ownership' in the Cloud," *Queen Mary University of London, School of Law Legal Studies Research Paper No. 45/2010*, 2010.

[5] Google Inc., "Ads in Gmail and your personal data " 2011; http://mail.google.com/support/bin/answer.py?answer=6603.

[6] A. Eichler, "Google Accidentally Resets 150,000 Gmail Accounts," 2011; http://www.theatlanticwire.com/technology/2011/02/google-accidentally-resets-150-000-gmail-accounts/20949/.

[7] Z. Zorz, "RSA hacked, SecurID users possibly affected," 2011; http://www.net-security.org/secworld.php?id=10763.

[8] D. Goodin, "User data stolen in Sony PlayStation Network hack attack," 2011; http://www.theregister.co.uk/2011/04/26/sony_playstation_network_security_breach/.

[9] G. Peev, "Fears over patient data as NHS computers are hacked into by 'pirate ninjas'," 2011; http://www.dailymail.co.uk/news/article-2001816/NHS-computers-hacked-Fears-patient-data.html#ixzz1PCj9KYIM.

[10] P. Thibodeau, "Amazon cloud outage was triggered by configuration error," 2011; http://www.computerworld.com/s/article/9216303/Amazon_cloud_outage_was_triggered_by_configuration_error.

[11] Z. Rezaee, A. Sharbatoghlie, R. Elam and P.L. McMickle, "Continuous auditing: Building automated auditing capability," *Auditing*, vol. 21, no. 1, 2002, pp. 147-164.

[12] G.H. Kim and E.H. Spafford, "The design and implementation of tripwire: A file system integrity checker," *Proc. 2nd ACM conference on computer and communications security (CCS '94)*, ACM, 1994, pp. 18-29.

[13] HyTrust, "HyTrust Appliance," 2010; http://www.hytrust.com/product/overview/.