JOURNAL OF INFORMATION SECURITY AND APPLICATIONS 27-28 (2016) 85–102

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jisa

IGOD: identification of geolocation of cloud datacenters

Chetan Jaiswal *, *Vijay Kumar*

Department of Computer Science and Electrical Engineering, University of Missouri, Kansas City, MO, USA

ARTICLE INFO

Article history:
Available online 2 February 2016

Keywords:
Data security
Cloud storage
Data storage policy
Data storage verification
Datacenter geographic location
Privacy

ABSTRACT

The geolocation of data has become a key concern with the evolution of cloud computing. Although data migration is quite common and sometimes essential for the purpose of load balancing or service guarantees, at times, it creates a risk for the user and could even violate the service agreement. A malicious service provider could also relocate the data, which could jeopardize data privacy and security. In this paper, we introduce a novel algorithm called IGOD to geolocate a cloud data center which can also be used for geolocating internet nodes. IGOD efficiently geolocates any target data center with higher accuracy and less cost. It provides audit control and assurance against such cloud storage providers who may move around a customer's data. We analyze and compare IGOD with currently available solutions of geolocating a target. We have used PlanetLab to validate IGOD and establish its cost-effective feature. To do so, we first use our own data collection to geolocate the test data center using emulation and compare IGOD's performance with other schemes. Finally, we use it to geolocate one of the Amazon S3 data centers. Our comparison shows that IGOD provides relatively higher location accuracy and is cost-effective (uses less resources).

1. Introduction

Cloud computing is becoming a globally accepted platform for data and activity management. The United States federal government announced its "Cloud First" policy and plans to migrate about 75% of their data management tasks on the cloud (Washington Post, April 17, 2011 and November 22, 2010) to comply with the policy. Many other business organizations are also renting or developing their own cloud platforms for migrating their data processing activities. For example, applications like Netflix, companies like Adobe, organizations like NASA and CIA use the AWS (Amazon Web Services) cloud for their data processing and data storage needs (AWS Case Studies: Government and Education). However, some organizations (mainly financial and medical ones such as banks, mortgage companies, hospitals, etc.) are reluctant to join the crowd for the right reasons; the most important being security, privacy, and trust (Groenfeldt; Vossler). Some might argue that few financial organizations have started using the cloud; however, it is only for the data analytics rather than storing customer information. Most of the financial organizations use a hybrid cloud for storing customer information and data analytics. For example *De Nederlandsche Bank* uses AWS for mobile applications, retail banking, high performance computing and credit risk analysis. A Forbes article (Groenfeldt) suggests "*The different demands enable banks to choose from several types of cloud applications such as private clouds, for the more sensitive data, and public clouds to store other information. More frequently, banks are going with a hybrid model that combines the two*". To safeguard data security and customer privacy, organizations tend to use private clouds for sensitive information like customer

* *Corresponding author.* Department of Computer Science and Electrical Engineering, University of Missouri, Kansas City, MO, USA. Tel.: +18164169750; fax: +1-(816) 235 5159.
E-mail address: chetanjaiswal@mail.umkc.edu (C. Jaiswal).
http://dx.doi.org/10.1016/j.jisa.2016.01.001

information (including SSNs) and EHRs (electronic heath records) and use public clouds for non-sensitive data and analysis. We observe that one of the privacy and trust concerns, which affects security is the location of data on the public cloud (Jansen, 2011; Kandukuri et al., 2009). A HIPAA (The Health Insurance Portability and Accountability Act) compliant Datacenter (DC) white paper also concurs with this concern related to the location of a DC and states that – "*Knowing where your data lives is a key consideration — if your data leaves the country, do you still have control of it? DCs operating outside of the country do not have to comply with HIPAA regulations, as HIPAA is created and enforced by the United States Department of Health and Human Services. Once your data travels overseas, it is possible you will be put at risk of a data breach or HIPAA violation, since international vendors are not required to observe our federal security regulations*" (HIPAA compliant data centers full access).

The locations of DCs are usually kept hidden by the service provider from the owners of the data. Some of the service providers do ask the customer preferred DC locations to store their data, however, there is no way for customers to verify or audit the provider's claim. Some organizations, such as financial ones, banks, etc. argue that if they do not know where their customers' data are located (which could be quite unsafe), they cannot guarantee their customers that their data is safe and their privacy is protected. Unfortunately, under the present model, location-aware or location-dependent DCs cannot be created without changing the cloud model. In the absence of these policies, one of the effective ways to identify the geographical location of DCs is to incorporate a location discovery facility at the application level. In this paper, we investigate this option and present our scheme, referred to as IGOD, to efficiently, accurately, and with less cost geolocates the desired DCs.

Another important issue is the enforcement of a data policy that defines the placement of DCs. Cloud providers must follow the service level agreement (SLA) with the customer and the regulations posted by the law enforcement agencies of each country. To reduce management costs, cloud providers may choose to store the data at a third party location that may not be in the geographical boundary of a country. In this case, the providers have to follow the security policies of the host countries. These issues are mainly governed by the policy and do not fall within the scope of this paper. However, the result of policy enforcement affects the location of the DCs, and for that reason, it relates to our investigation.

2. Our contribution

The location discovery of DCs is a relatively new issue that emerged with the widespread use of cloud computing. In one of our earlier papers on security (Jaiswal et al., 2014), we argued that knowing the location of DCs is quite helpful in developing location-dependent security schemes. So far, there is no indication that cloud providers are interested in disclosing the location information.

This has motivated researchers (including us) to develop application level schemes for geolocalization of a DC where the desired file is stored. Recently, there has been a number of reports presenting schemes for geolocating DCs. These schemes work fine; however, they are prohibitively expensive and comparatively less accurate with a high error margin. Furthermore, most of these works are IP-based and geolocate network endpoints. Unfortunately, they cannot be used to geolocate a DC in STaaS (Storage as a Service model, such as Amazon S3) (AWS | Amazon Simple Storage Service; Building Storage-as-a-Service Businesses) because of the absence of an IP address or any other identifying information about the DC. These are the motivating factors, and since the need for this information will grow with expansion of the cloud system, we have developed a general algorithm IGOD, that geolocates the DC where the user's data are stored. IGOD overcomes the limitations of earlier works and our results demonstrate that it offers a relatively higher location accuracy and uses fewer LMs and thus, significantly lowers the geolocating cost. One of the additional benefits of IGOD is that it provides a non-repudiation service that gives users a tool for audit control by providing strong assurances. In this work we have assumed that the data of a user is not partitioned and thus it is not stored across multiple DCs. This is a reasonable assumption and concurs with the previous work (Gondree and Peterson, 2013). Nevertheless, this data can be replicated in its entirety on multiple DCs. IGOD is capable of identifying the service point from where the data (File-1) is being served. A service point is the communication control server of a DC which entertains the requests of the users and routes it inside the DC to an appropriate node. There can be replicated copies of the File-1 at different locations around the globe at different DCs; some being active copies and others as passive. By active copy we mean that this copy will be served in response to a query from the user. A passive copy is just a backup file and would be used in case of disaster recovery. By definition, if we have one or more active copies; there can be one or more service points (or DCs) and IGOD is capable of identifying them. This however does not mean that IGOD will find the location of all the copies (passive) of the data stored in the cloud as it is not necessary and feasible to find the passive copies of the data in the cloud. Should they become active, another run of IGOD will identify their location as well. We have presented our argument in Section 5.4.

This paper is structured as follows: Section 3 provides a review of earlier works; Section 4 explains our approach and algorithm; Section 5 provides the evaluation, and the paper finishes with its conclusion in Section 6, and acknowledgements and references after.

3. Review of earlier works

There are two categories of schemes that have been reported in the literature (a) locating web servers or Internet hosts (Gill et al., 2010; Gueye et al., 2006; Katz-Bassett et al., 2006; Laki et al., 2011; Padmanabhan and Subramanian, 2001; Percacci and Vespignani, 2003; Watson et al., 2012; Youn et al., 2009) and (b) geolocating DCs (Albeshri et al., 2012; Benson et al., 2011; Fotouhi et al., 2015; Gondree and Peterson, 2013). To the best of our knowledge, out of these, only one report has presented a scheme for geolocating DCs in STaaS model. One can argue that (a) L/L (Longitude/Latitude) from DNS or (b) Whois