Current Security Threats and Prevention Measures Relating to Cloud Services, Hadoop Concurrent Processing, and Big Data

Ather Sharif, Sarah Cooney, Shengqi Gong, Drew Vitek
Department of Computer Science
Saint Joseph's University
Philadelphia, PA 19131
{ather.sharif, sc599170, sg632463, av584724}@sju.edu

Abstract—**Cloud services are widely used across the globe to store and analyze Big Data. These days it seems the news is full of stories about security breaches to these services, resulting in the exposure of huge amounts of private data. This paper studies the current security threats to Cloud Services, Big Data, and Hadoop. The paper analyzes a newly proposed Big Data security system based on the EnCoRe system which uses sticky policies, and the existing security architectures of Verizon and Twilio, presenting the preventive measures taken by these firms to minimize security concerns.**

Keywords: Cloud Services, Big Data, Hadoop, Security

I. Introduction

In a 2008 paper by three members of the Computing Community Consortium, Big Data was declared the biggest innovation in computing of the past decade [1], and soon became a term familiar even to laymen as its potential was recognized in fields such as healthcare, public sector administration, retail, manufacturing, and personal location data, to name a few [2]. With the increase in popularity of social media and the introduction of open APIs, the need to store and analyze thousands of terabytes of data became an even more prominent and pressing challenge in Computer Science.

The year 2009 is considered a milestone year in the field of Computer Science. In this year, Web 2.0 reached a significant level of popularity and inspired technology giants such as Microsoft and Google to offer browser based enterprise solutions. Although, Amazon had actually released EC2/S3, considered to be the first cloud computing service, in 2006 [3], 2009 marked the takeoff of cloud technology.

In December 2011, Hadoop (originally developed in 2005 as a support distribution for the Nutch search engine project at Yahoo!) released Release 1.0. This release was the first with security support, a seemingly obvious and imperative feature that was not part of previous releases. Hadoop has since emerged as one of the most cutting-edge technologies to store, process and analyze Big Data through use of a cluster scale-out environment. Hadoop is widely used worldwide by a variety of large companies including Adobe, Facebook, and Spotify. According to the International Data Corporation, it is expected to grow its already significant software market to $812.8 million in 2016 [4].

Cloud based computing services offering storage, infrastructure, platform, and software services have become widely available across the commercial, noncommercial and academic sectors over the past several years. Many of these systems have been especially targeted toward Big Data collection and analysis, for instance, platforms for allowing Hadoop to run in the cloud environment. These services are cost-efficient, highly scalable, and are offered by many prominent companies such as Microsoft, Google and Amazon. As usage of such services has increased, their security has been given a great deal of attention in various forums and discussions over the past few years. Although the services offered by prominent companies are backed by the reliability of the company name, they too have been called into question over their security provisions. This paper discusses the current security threats that cloud based services for Big Data storage and analysis, including Hadoop, are facing and the preventive measures deployed in response to such threats. Sections two, three, and four focus on security architectures and preventive measures for Cloud Services, Big Data and Hadoop, respectively. Section five concludes the paper.

II. Cloud Services

According to an article published by Forbes in 2013, at that time more than half of the businesses in the United States were using cloud services [5], and this number has only continued to grow. As this technology matured, many data, security, and privacy concerns, for instance the possibility of data theft, were identified and had to be addressed. However, a recent report shows that overall user confidence in the security measures taken by cloud service providers is still very low [6]. Customers have concerns about the basic Cloud principle of shared resources, the lack of control over where their data is being stored, and the lack of openness by providers about what security measures they are taking [6].

In principle, Cloud services are dependent on the technology environment of the service provider. Thus, it is up to the service provider to address many of the questions and concerns pertaining to security. The 'Notorious Nine Cloud Computing Threats" identified by the Cloud Security Alliance as the most pressing threats to Cloud security, and their relevance in the present era, calculated from a poll of industry experts by the Cloud Security Alliance, are displayed in Table I [7].

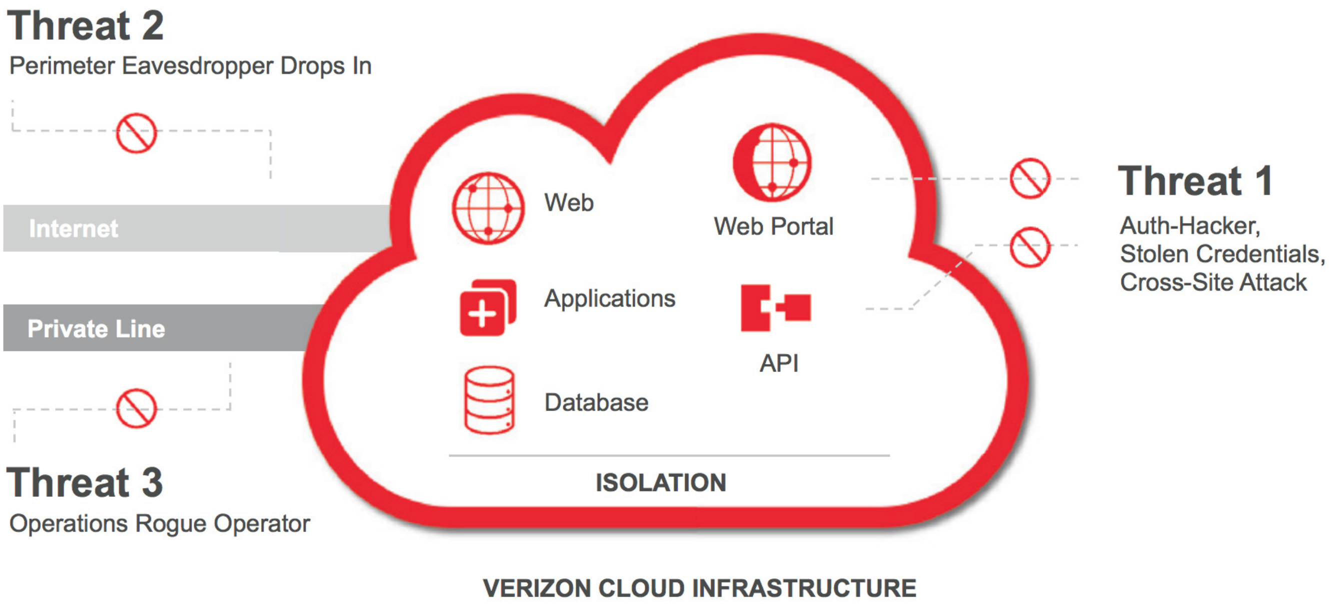

Fig. 1. Verizon Cloud Infrastructure.

This paper studies Verizon's infrastructure, as shown in Figure 1. Verizon uses a layered security model to secure its Cloud [8], a service provided to its customers to store the personal data and information shared across their devices. The four layers are Base Security, Logical Security, Value-Added Security, and Governance, Risk and Compliance.

A. Base Security

This layer is focused on physical, or external, security independent of the technology itself. Verizon has made efforts to ensure that their data centers have the highest level of physical security, utilizing measures such as around the clock video monitoring with 90 day video support, biometric readers, constant on-site guard services, and in-bound shipment security that only accepts pre-notified packages. Additionally, access control is maintained via ”need to know” and ”event by event” protocols. These measures help mitigate the risks of physical data theft and data loss via physical attack of hardware. Employees receive regular training on up-to-date security procedures. All employees must also pass background checks to diminish the risk of malicious insider attacks.

TABLE I. SECURITY THREATS WITH THEIR RELEVANCE.

Security Threats	Relevance
Abuse of Cloud Services	84%
Account or Service Traffic Hijacking	87%
Data Breaches	91%
Data Loss	91%
Denial of Service	81%
Insecure Interfaces and APIs	90%
Insufficient Due Diligence	81%
Malicious Insiders	88%
Shared Technology Vulnerabilities	82%

B. Logical Security

The Logical Security level ensures the confidentiality, integrity, and availability of networks, resources and data. This level is further divided into Compute, Network, Storage, and Management sublayers that work together to protect the infrastructure. In the Compute sublayer, measures such as password policies, operating system security, and authentication of administrators and users are implemented. The Network sublayer accounts for Distributed Denial of Service (DDoS) detection and mitigation, ”MAC-in-MAC” encapsulation at the firmware level, and integrated firewall capabilities. The Storage sublayer supports the encryption of data using a symmetric AES-256 cipher and the further enhancement of security through SSL and advanced sanitization techniques. This layer also gives users the option to add database encryption through Microsoft SQL and Oracle. Finally, the Management sublayer handles identity and access control. Verizon hopes to support the Security Assertion Markup Language (SAML) 2.0 Framework in its next release.

C. Value-Added Security

This layer encapsulates some additional security features supported by Verizon besides Base and Logical security. These features include Firewall and VPN capabilities with highly customizable structures, preconfigured security solutions that include Big Data applications and software development features, and an intelligent management system, which is capable of detecting security vulnerabilities and identifying mitigation options. The Value-Added Security layer also provides an option for a Private IP network that supports an end-to-end environment for secure connectivity to the users' workloads.

D. Governance, Risk and Compliance

This layer ensures all security measures in the three previous layers comply with well-established and reputed standards. Updates, enhancements, and bug fixes are done using agile development techniques. By the nature of agile development, fixes are performed with strong controls, and rapid innovation is ensured. High-end change management is also used during updates and debugging to support verification, rollback procedures, and prerequisites.

III. Big Data

With social media more popular than ever before and the growing amount of data in numerous other sectors, Big Data analysis has become increasingly important to both large companies and small businesses.

The analysis of Big Data provides helpful business insights that contribute toward growth, rectification of unproductive practices, and the highlight of strengths and weaknesses, among other benefits. These days, BigData is increasingly stored and analyzed using Cloud services, and while the collection and analysis of Big Data can have a highly positive business impact, it can also prove to be extremely detrimental without proper security measures. This paper studies the security concerns for Big Data in terms of three V's, namely volume, variety, and velocity [9].

The sheer volume of data being stored and analyzed presents a challenging task for encryption. The heavy cost of encrypting and decrypting large volumes of data affects the overall speed of viewing and using the data. However, encryption is necessary to avoid the co-mingling of datasets within the Cloud framework and to avoid vulnerabilities that could lead to data theft or loss. Once the data is encrypted and stored, analysis should be performed without decrypting the data to ensure security and privacy of the client [10].

Different types of data in a single dataset may need to be accessed and or used differently, and handling the access control for the variety of data types is a difficult operation. Considering the fast velocity of the data and that there is no fixed path of travel between nodes, tracking access and data flow can be challenging, and present a security concerns. However, once again if access and data flow are not properly controlled, the data can be left vulnerable to theft and loss.

Without proper data flow control, there are also risks associated with the shared technology environment essential to Cloud services. If a corruption in one data set, maliciously intended or not, is allowed to move outside the proper bounds, there is a risk of corrupting other customer data sets or, worst case, the Cloud service provider's entire infrastructure. The large volume and fast velocity of data can also cause the backup and restore functions to adversely affect performance, but leaving them out is not an option if data loss is to be avoided.

In the next section, this paper studies the architecture of a newly proposed sticky policy framework for securing Big Data applications.

A. Sticky Policies

Li et al. [11] propose a framework with a unique architecture to aid in securing Big Data applications, based on the data privacy management project EnCoRe [12]. EnCoRe implements sticky policies, which apply constraints and conditions on data in order to define usage allowances and obligations.

This controls the access to and disclosure of confidential data across several boundaries within a project. EnCoRe uses public-key-encryption as its core mechanism to manage sticky policies.

The Big Data framework proposed by Li and his co-authors uses loose-couple binding, a method that stores the sticky policies and data fragments in separate places. This increases compliance in the infrastructure and makes the system harder to breach. The architecture has two sub-domains, the trusted authority domain and the data center domain, for the purpose of keeping the sticky policies and data fragments from comingling. The trusted authority domain holds the identity and key management engine and policy engine, while the encrypted data is housed within the data center domain.

Information about users including their authentication and authorization information, and the privileges a given user holds for each piece of data is stored within the identity and key management engine. The policy engine functions as the core of the entire domain. This engine maintains control of the data being accessed and keeps track of parties' privileges in regard to accessing the data. On top of this, the stored data is encrypted and can only be accessed after policies have been approved, accepted, and satisfied.

The policy engine has several sub-components: the policy portal, the policy controller, the policy negotiation component, the policy update component, the enforcement component, and the policy store. The policy portal serves as the entryway to the engine, receiving requests for data access, and sending final responses back to the user. The policy controller makes a decision to either reject the request or forward it to the responding component. True to its name, the policy negotiation component negotiates policies, matching policies to a database of those users with authority to carry it out. Security policies can be updated through the policy update component. The enforcement component checks whether or not the data user has fulfilled the required sticky policies. Finally, the policy store keeps a, mapping between data and the policies required to access it, as well as maintaining audit logs of all the activated policies and accessed data. To fulfill its functions, the policy store sustains three subdatabases, the policy database, the sticky database, and the audit database. Simply put, the policy database stores the policy rules, the sticky database stores the mapping between the sticky policies and the data to which they are "stuck", and the audit database logs and tracks data access. The framework provides a theoretically thorough design to yield more secure transactions within the realm of collecting, analyzing, storing, and sharing Big Data. This framework could be especially relevant for users who work with private of classified data like hospitals storing medical records or law enforcement groups analyzing top secret investigation reports.

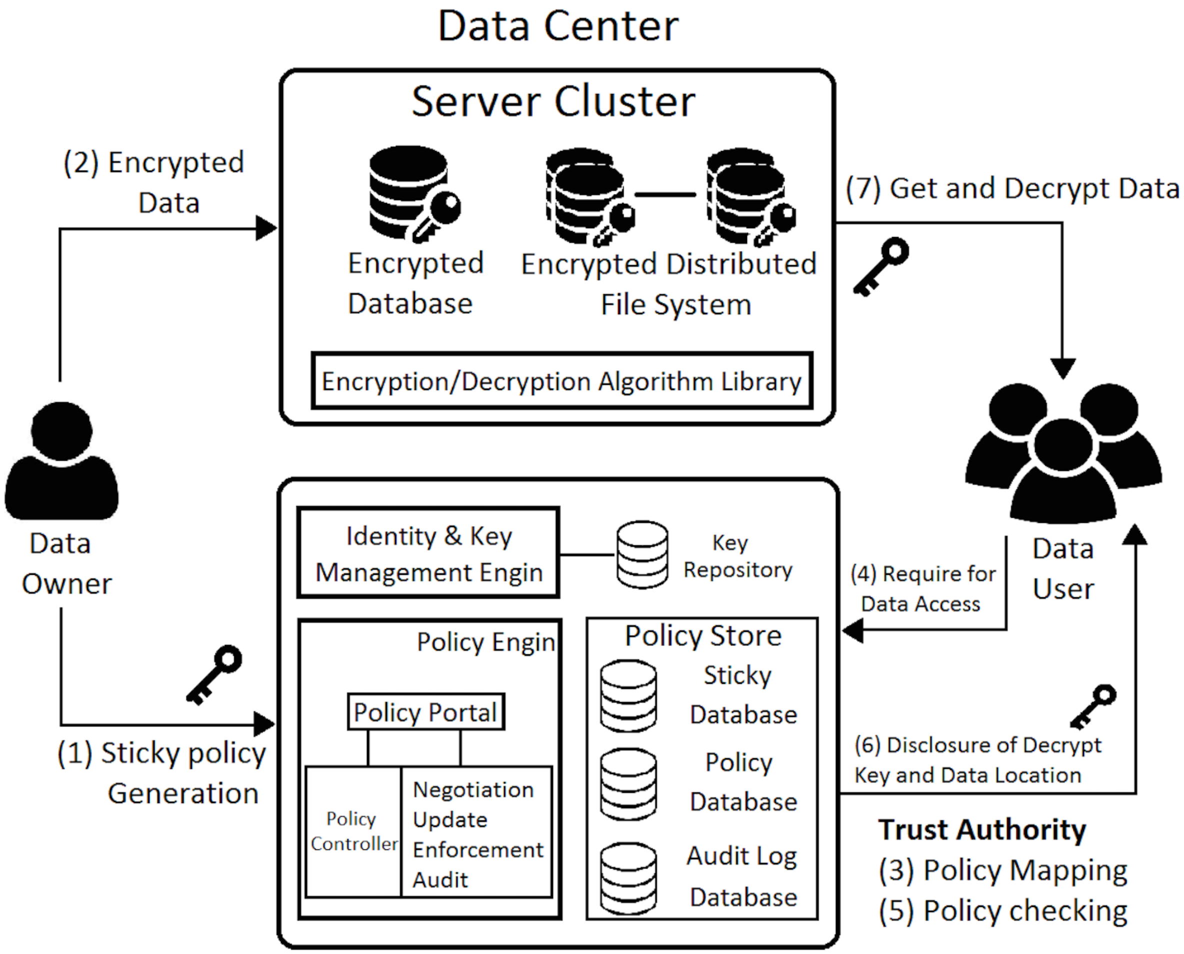

Fig. 2. Sticky Policy Framework for Securing Big Data Applications.

IV. Hadoop

Hadoop's open-source framework for Big Data analysis has been gaining popularity since its inception. With its popularity increase, Hadoop's security has continually come under scrutiny, especially given its concurrent processing architecture. In response to authentication vulnerabilities [13], Hadoop introduced central authentication through Kerberos. In 2013, ”Project Rhino” was introduced by Intel, which strengthened the security of the Hadoop framework by providing support for a variety of encryption and authentication methods. However, based on the massive increase in data volume and critical information being transmitted in the data lake, there still exists a need to strengthen the security measures of the Hadoop framework in an effort to prevent cyber-crime [11]. The Hadoop framework supports traditional distributed systems as well as systems that are part of a cloud environment. This paper focuses on the security vulnerabilities and preventive measures for systems connected through a cloud. The architecture of Hadoop allows a file to be broken into chunks, which are distributed across various nodes to be processed concurrently. However, in the cloud environment it is very difficult to locate the specific node holding a given chunk of data. One major security threat in this system regards the protection of critical data, which needs extra security measures to be in place. As mentioned in section II, the variety of security levels in a single data set complicates security procedures. The Hadoop chunk system only compounds this problem. Further, a node can contain chunks from multiple files and has administrative rights. ”Chunk stealing” and ”chunk injection”, taking or modifying data without proper permissions, are two other major security issues.

This paper studies the security architecture of Twilio, a cloud communication company that is well-known for its reliable implementation of the Hadoop framework using Amazon S3 services [14]. Twilio uses a multi-tenant communications platform to ensure the isolation of rights to resources, as shown in Figure 3. This strategy is further extended to identify critical data and perform necessary security checks. In order

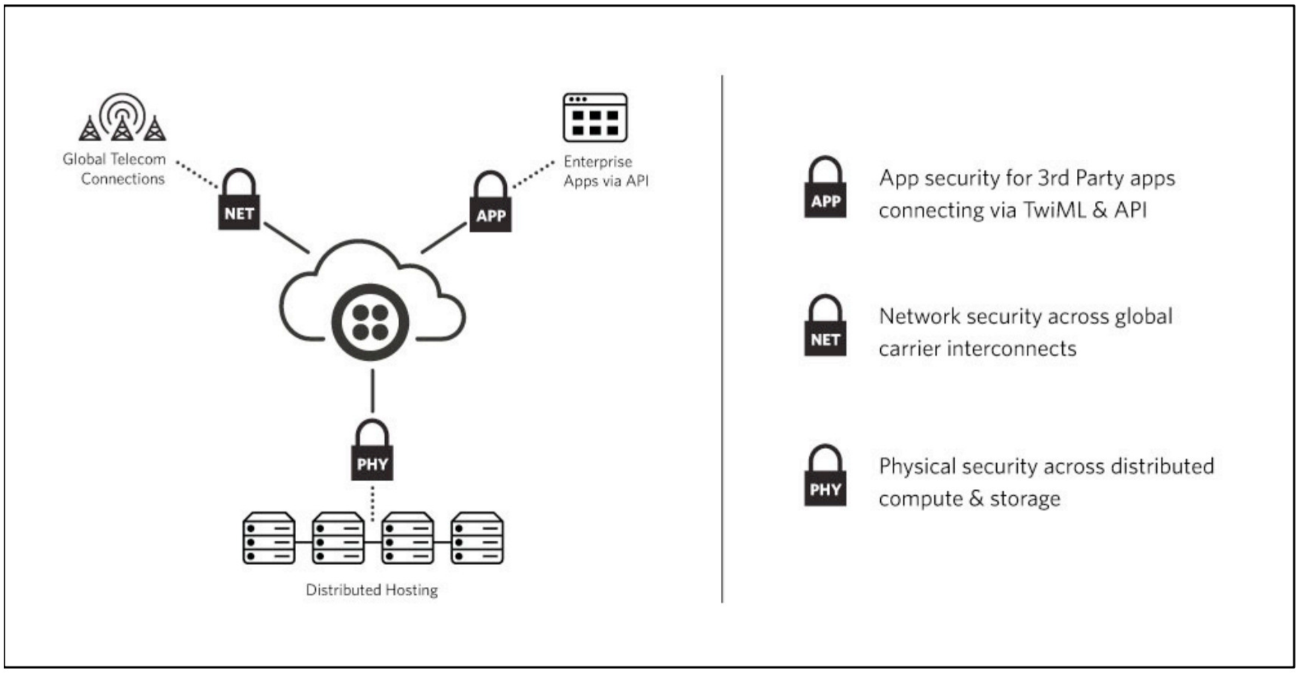
Global Telecom Connections
NET
APP
Enterprise Apps via API
PHY
Distributed Hosting
APP
App security for 3rd Party apps connecting via TwiML & API
NET
Network security across global carrier interconnects
PHY
Physical security across distributed compute & storage

Fig. 3. Twilio Cloud Security Architecture.

to avoid ”chunk stealing”, Twilio uses access controls based on job roles combined with Amazon S3’s bucket policies and Access Control Lists. Twilio’s compliance with Safe Harbor standards contributes to a high level of trust by third-party vendors, many of whom are required to have separate standard security compliances. Further attacks such as Denial of Service and Distributed Denial of Service are minimized through maintaining redundant DNS servers and blocking rogue IPs using iptables, as well as through supporting AWS proprietary DDoS mitigation techniques. Physical theft of data is still a possibility, but is protected against using base security measures similar to those used by Verizon in securing its Cloud, as described in section II A, as well as multiple redundancy zones in several different geographic regions. The overall security is assessed every six months via penetration testing by third party companies.

V. Conclusion

Big Data collection and analysis and cloud computing services like Hadoop are being utilized on an increasingly wide and global scale. The massive increase in the amount of data produced even in just the last several years also means an increasing number of technologies are being developed to collect, analyze, and store this data, each with their own new vulnerabilities and security threats. The massive number of data breaches across entities from retail stores to government databases in the news today is proof that security cannot be an afterthought in developing these Big Data systems. Today, there is continuous work going into identifying security concerns in existing systems and developing measures to minimize or eradicate them. As new systems are developed it is imperative that security be made a top priority in the initial design and implementation. Although, even if security is prioritized, new vulnerabilities are sure to be identified. Furthermore, a unified system to completely stop cyber criminals has yet to be designed, although the implementation of theoretical systems like the sticky-policy architecture may hold promising improvements. The Big Data security field is certain to continue to play an imperative role in the technical world as Cloud services increase in popularity for everything from providing cellular customers with personal data backup to large scale storage and analysis of Big Data by corporations large and small.

VI. Acknowledgements

The authors of this paper would like to thank Dr. Babak Forouraghi, Professor, Chair and Graduate Program Director of the Department of Computer Science at Saint Joseph’s University (SJU) for his continued support, guidance, and the knowledge conveyed through the Introduction of Security course at SJU.

References

[1] R. Bryant, R. H. Katz, and E. D. Lazowska, “Big-data computing: creating revolutionary breakthroughs in commerce, science and society,” December 2008.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, and M. G. Institute, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute San Francisco, 2011.

[3] A. Mohamed, “A history of cloud computing,” March 2009. [Online]. Available: http://www.computerweekly.com/feature/A-history-of-cloud-computing

[4] I. D. Corporation, “Idc releases first worldwide hadoop-mapreduce ecosystem software forecast, strong growth will continue to accelerate as talent and tools develop [press release],” May 2012.

[5] R. Cohen, “The cloud hits the mainstream: More than half of u.s. businesses now use cloud computing,” http://www.forbes.com/sites/reuvencohen/2013/04/16/the-cloud-hits-the-mainstream-more-than-half-of-u-s-businesses-now-use-cloud-computing/, April 2013.

[6] J. Bourne, “Security failing to keep pace with cloud technology adoption, report finds,” http://www.cloudcomputing-news.net/news/2015/apr/09/security-failing-keep-pace-cloud-technology-adoption-report-finds/, April 2015.

[7] C. Alliance, “The notorious nine: Cloud computing top threats in 2013,” *Cloud Security Alliance, Tech. Rep*, 2013.

[8] Verizon, “What makes a cloud secure,” February 2015.

[9] R. Sowmyanarayanan, “Securing big data,” http://tdwi.org/articles/2015/04/14/securing-big-data-pt1.aspx, April 2015.

[10] V. O. Waziri, J. K. Alhassan, I. Ismaila, and M. N. Dogonyaro, “Big data analytics and data security in the cloud via fully homomorphic encryption,” *International Journal of Computer, Control, Quantum and Information Engineering*, vol. 9, no. 3, 2015.

[11] S. Li, T. Zhang, J. Gao, and Y. Park, “A sticky policy framework for big data security,” 2015.

[12] S. Pearson and M. C. Mont, “Sticky policies: an approach for managing privacy across multiple parties,” *Computer*, no. 9, pp. 60–68, 2011.

[13] O. O’Malley, K. Zhang, S. Radia, R. Marti, and C. Harrell, “Hadoop security design,” *Yahoo, Inc., Tech. Rep*, 2009.

[14] T. C. Communications, “Security architecture,” July 2014.