ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

Detection of Malicious Attackers in Big Data System using Data Replication Techniques

Sukeshini Gawai[1], Madhuri Bidwe[2]

Lecturer, Department of Computer Engineering, D Y Patil Polytechnic, Akurdi, Pune, Maharashtra, India[1]

Lecturer, Department of Computer Engineering, D Y Patil Polytechnic, Akurdi, Pune, Maharashtra, India[2]

ABSTRACT: Infrastructureof big data system is arranged in the large amounts of data are hosted away from the users. In such a system information security is considered as a major challenge. From a customer perspective, one of the big risks in adopting big data systemsis in trusting the provider who designs and owns the infrastructure from accessing user data. Yet there does not exist much in theliterature on detection of insider attacks. In this work, we propose a new system architecture in which insider attacks can be detectedby utilizing the replication of data on various nodes in the system. The proposed system uses a two-step attack detection algorithm anda secure communication protocol to analyze processes executing in the system. The first step involves the construction of controlinstruction sequences for each process in the system. The second step involves the matching of these instruction sequences amongthe replica nodes. Initial experiments on real-world Hadoop and spark tests show that the proposed system needs to consider only 20%of the code to analyze a program and incurs 3.28% time overhead. The proposed security system can be implemented and built for anybig data system due to its extrinsic workflow.

KEYWORDS: Big Data, Security, Intrusion Detection, Insider Attacks, Control Flow.

I. INTRODUCTION

Traditional security methods are necessary but not sufficient for big data systems. Big data security has some unique challenges concerning both applications and data. For instance, current big data security platforms focus on providing fine-grained security through extensive analysis of stored data. But such models indirectly facilitate the abuse of user data in the hands of applications and service provider. This led to the rise of differential privacy that aims at protecting sensitive user information while supporting data analytics. Another such security concern that has been seldom addressed in the big data world is insider attacks. Insider attacks are becoming more common and are considered the toughest attacks to detect. There does not exist much in the literature on solutions for insider attacks in general [3]. Existing insider detection techniques concentrate on user profiling and access control. For these methods to be applicable in the big data world, it is assumed that collusion is a rare event. Though the assumption holds true in most cases, the real drawback with existing insider detection techniques is their inability to be applied in distributed compute environments. To the best of our knowledge, there is no robust solution for detecting or preventing insider threats within big data infrastructures. But it is crucial to address the problem of insider attacks in big data systems for three main reasons: (a) traitor within the provider's organization will be able to circumvent the security system in place (b) sensitivity of customer information stored in the system is increasing by day; and (c) there is no consensus or widespread agreement on well-defined security standards in the big data community.

Big data solutions are widely adopted across various government and enterprise domains such as software,
finance, retail and health care. They are pioneering in the field of advanced data analytics and have a projected market of approximately 50 billion dollars by 2018. The most frequent use-cases of big data are information retrieval from complex, unstructured data; and real time data analysis. But along with its rapid market growth, the big data trend also has its share of challenges and risks. In an era where extracting information from data is sanctioned to all, users are understandably more skeptical to let service providers host their data away from them. This, along with the recent increase in the number of cyber-attacks elevates the importance for security in big data. Though privacy and security are touted to be important problems in the big data world, the solutions concentrate only on leveraging big data systems

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

for efficient security in other domains. According to Open SOC, in 60% of breaches data gets stolen within hours of the breach and 54% of breaches are not discovered for months [9]. This is unacceptable in the big data world where the revenue is based on efficient management of user data. Recently, two unauthorized backdoors were discovered in Juniper Networks firewalls that might have given attackers access to highly classified information. Some important facts about this particular hack are: (a) it comes at the cost of compromising national security (b) it shows that even a major network security company is vulnerable to attacks (c) it is believed that these backdoors were left undiscovered for almost 3 years; and (d) it was reported that the attackers could have deleted the security logs [8]. To fight the evolving scope of attacks and attackers, it is important to apply traditional security methods in new

II. RELATED WORK

Insider Attacks:

Though security in general computing has been extensively studied and implemented over the years, computers are still vulnerable to attacks. Software based attacks that typically target a computer network or system, called cyber-attacks, are growing in their frequency and impact. The plot for any type of software attack involves exploitation of a piece of code that runs on a computer. It is inherent to this perspective about a cyber-attack that security can be provided at two levels: (a) by the software that is used to compile and execute the program; and (b) by the hardware that runs the program. Providing security at software level gives more context and information about the target programs that are being protected. But this comes with the risk of the security software itself being compromised.
On the other hand, having security at hardware level gives more isolation to the process of analyzing and securing programs though it becomes difficult to give detailed context about the programs and the infrastructures running them. In any case, the toughest software attacks to counter are the ones whose genesis is intentional and are performed by those who have a good understanding of the underlying system. Based on our literature review, we have identified four major questions that can guide towards better handling of insider attacks: (a) who can perform these attacks? (b) what gets affected? (c) how to detect these attacks? and (d) how to prevent them from happening? Figure 1 gives a list of entities to consider when dealing with insider attacks. The figure also shows the four questions, from above, as relationships among the entities. Insider attacks can be performed by (a) traitors who are legally a part of the system but want to misuse the access privileges given to them; (b) masqueraders who get access to the system by stealing identities of those who have legitimate access. Insider attacks can affect the proper functionality of a program or corrupt the data used by the programs.

Insider Attacks in Big Data Systems

Insider attacks are a dangerous security problem in any domain because they are difficult to predict and detect.Hence organizations must try to safe guard their systems and data from insider attacks. Predictive models foruser/program/network behavior with the help of continuous monitoring is a widely adopted solution for insiderattack detection. But such prediction is not completely reliable and the difficulty in detecting attacks grows with thecomplexity of the underlying system. Recent advancements in computing led to wide adoption of services such ascloud computing and big data which are extremely complex in their design and development. In cloud computing,many insider attacks can be performed by misleading the client side services and once compromised, data obtained can provide social engineering opportunities for cascade attacks. Having security as a service model for cloud environments and having sealed clouds are some ideas proposed towards protecting cloud infrastructures from insider attacks. While cloud computing is more about computing on the fly, big data deals with organizing and managing large sets of data. Insider attack detection and prevention for big data frameworks is an area that is not well explored yet. Security within big data systems is still a budding phenomenon.

III. PROPOSED SYSTEM

In the proposed system we include a secure communication protocol and a two-step attack detection algorithm. The first step in the attack detection algorithm isprocess profiling, which is conducted locally and independentlyat each

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

node to identify possible attacks. In the nextstep is hash matching and consensus, which is conductedby replica data nodes to conclude about the authenticity ofa possible attack.

Secure Communication Protocol

A big data system is technically a distributed data storage system that relies on secure and efficient communication protocols for data transfer. The proposed system aims to provide robust security for big data systems by having a modular design and being independent from the core big data services. For this reason, a separate secure communication protocol is included in the proposed system design that can be isolated from the set of default communication protocols used by the big data system. The proposed system is a mix of independent security modules that work together and reside on individual nodes of the system. These modules use the secure communication protocol to share packets of data with their counterparts on other nodes of the cluster.

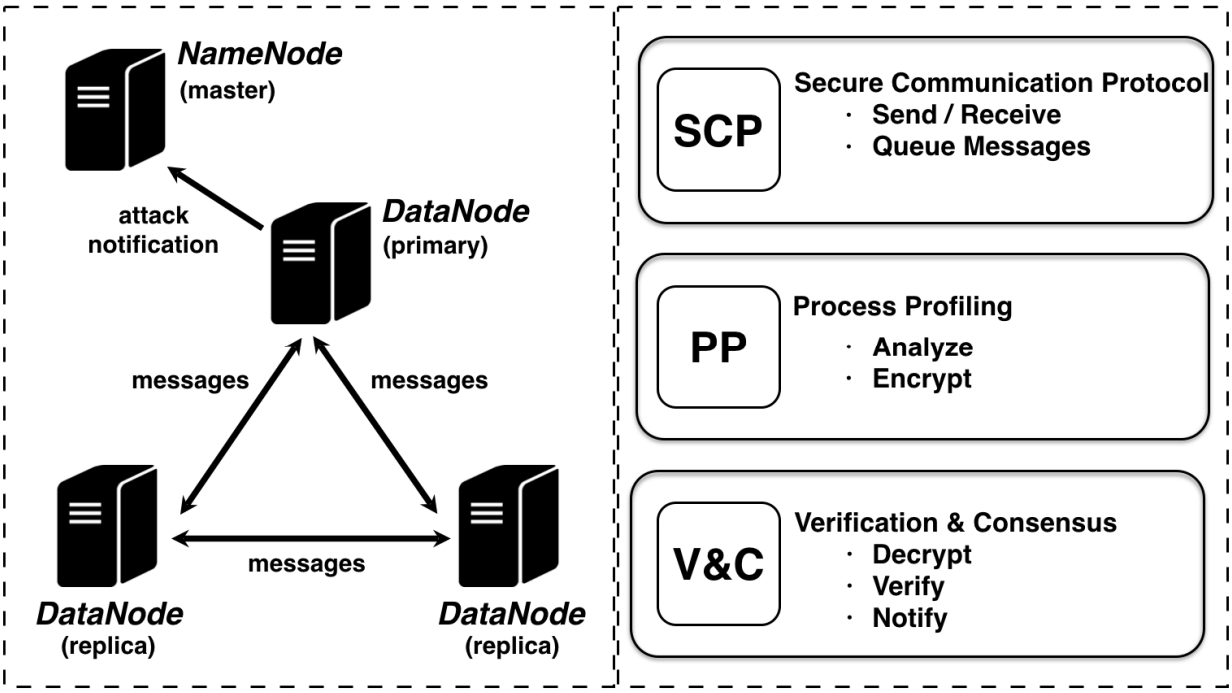

Fig. 1. Proposed System Architecture for Detecting Insider Attacks in Big Data Systems

Detection Algorithm

The main part of the proposed system is the attack detection algorithm which will be explained in this subsection. Our attack detection algorithm is a two-step process:
process profiling (step 1) and consensus through hash matching (step 2).

Step 1: Process Profiling
Traditionally vulnerability scanning is performed awayfrom the source program's execution domain to guarantee isolation. Hence, the results of such scan must be communicatedback to the program. But this leads to a cost versusisolation trade-off, depending on the remoteness of thelocation used to perform the vulnerability scan. In big dataapplications, the source program's execution is distributedacross multiple nodes of the cluster. This makes it difficultto implement techniques such as vulnerability scans on big data systems. But big data infrastructures use

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

replication ofdata for high availability. This enforces the same program to be run on multiple nodes that host the data required forthe program. We exploit this unique property of big datasystems and introduce a variation of CFI to create a novelprocess profiling technique that can help in detecting insiderattacks in big data systems. Evans et al.show that CFI,either with limited number of tags or unlimited number of tags, is not completely effective in attack prevention. Also, CFI is usually based on CFG created from static analysis of program code.

Algorithm

Hash Matching

while true **do**
msgpget message about process p from main copy
hashhashes(receivedp) decrypt(msgnew, privk)
hashhashes(localp) process − profile(p)
if hashhashes(receivedp) 2 hashhashes(localp) **then**
confirmation safe
else
confirmation unsafe
end if
send (confirmation, main)
end while

Model of the Proposed System Architecture

The proposed security system is a combination of 3 parts: secure communication protocol, process profiling and hash matching. As shown in Figure 1, these three parts are made of multiple modules that need to be installed on all nodes in the big data system. Also, locality of these modules impacts the performance of the system greatly. The closer they are to the main processor of a node, the faster and less expensive it will be to communicate. But from a security standpoint, these modules need to be isolated from the big data system main workflow. Hence we designed a model for the proposed system that can fit on isolated special purpose security hardware chips. Such chips can be built on top of existing security hardware such as TPM or Intel's TXT chips. Hardware solutions are popularly known to affect the scalability and flexibility of the big data infrastructure, comparing to a software solution which can be very adaptive. But in this case, we avoid such problems by decoupling our solution from the workflow of a big data platform.

IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed a security system for big data systems to detect insider attacks quickly with low overhead. The system consists of a two-step attack detection algorithm and a secure communication protocol. A simple hash string matching technique is proposed to fulfill the distributed process similarity check and identify attacks. A secure communication protocol for data nodes that uses periodically generated random keys is proposed to conduct the detection algorithm. A model of the proposed system is tested in real-time on Amazon's EC2 clusters using a different sets of Hadoop and Spark programs. The time overhead was 3.28% and it is observed from the results that the proposed security system uses only 20% of program code to detect attacks. In this work, we also propose the idea of delegating security as an independent module and the componentsneeded for such models are discussed. For future work, we would like to evaluate our system on security related big data benchmarks (when available). Also, we would like to actualize the hardware architecture of security chips that can independently support our system.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 8, August 2018

REFERENCES

[1] IDC. New IDC Forecast Sees Worldwide Big Data Technology and Services Market Growing to $ 48.6 Billion in 2019, Driven by WideAdoption Across Industries. IDC, 09 Nov. 2015. Web. 01 Jan. 2016.
[2] Vormetric. ”2015 Insider Threat Report.” Vormetric, Inc, 01 Sept. 2015. Web. 01 Jan. 2016.
[3] Salem, Malek Ben, ShlomoHershkop, and Salvatore J. Stolfo. ”Asurvey of insider attack detection research.” Insider Attack andCyber Security. Springer US, 2008. 69-90.
[4] White, Tom. Hadoop: The definitive guide. ” O’Reilly Media, Inc.”, 2012.
[5] Zaharia, Matei, et al. ”Spark: cluster computing with working sets.”Proceedings of the 2nd USENIX conference on Hot topics in cloudcomputing. Vol. 10. 2010.
[6] Neuman, B. Clifford, and Theodore Ts’ O. ”Kerberos: An authentication service for computer networks.” Communications Magazine,IEEE 32.9 (1994): 33-38.
[7] Aditham, Santosh, and NagarajanRanganathan. ”A novel framework for mitigating insider attacks in big data systems.” Big Data(Big Data), 2015 IEEE International Conference on. IEEE, 2015.
[8] Khandelwal, Swati. ”Juniper Firewalls with ScreenOSBackdooredSince 2012.” The Hacker News. The Hacker News, 18 Dec. 2015.Web. 01 Jan. 2016.
[9] Sirota, James, and SheetalDolas. ”OpenSOC.” Open Security Operations Center. Cisco, 05 June 2014. Web. 01 Jan. 2016.
[10] Bajikar, Sundeep. ”Trusted platform module (tpm) based security on notebook pcs-white paper.” Mobile Platforms Group Intel Corporation(2002): 1-20.
[11] Greene, James. ”Intel trusted execution technology.” Intel Technology White Paper (2012).
[12] Schultz, E. Eugene. ”A framework for understanding and predictinginsider attacks.” Computers & Security 21.6 (2002): 526-531.art. 1995 (1995): 4-5.