International Journal of Technology in Computer Science & Engineering

ISSN 2349 – 1582

Volume 4, No 2, June 2017

Virus Threat identification in Big Data systems

S.Natarajan[1]
(Research Scholar,Manonmaniam Sundaranar University,Tirunelveli,TN,India
sivnatarajan@gmail.com)
Dr.H.Abdul Rauf [2]
(Professor & Dean Sree Sastha Institute of Engineering and Technology, Chennai,TN,India
harauf@gmail.com)
Dr.S.P.Victor[3]
(Dean of Science and Associate Professor,St.Xavier's College,Palayamkottai,TN ,Indria)

Abstract: Computer viruses and worms have been becoming important security threats, over several decades. At present in any organization it is a challenging task to identify and rectify the problems caused by the viruses. Big Data systems provide more facilities to store huge volumes of data from various heterogeneous sources. There are lot of advantages of using Big Data systems for storing and retrieving our valuable data and files. If a part of these data is infected/corrupted by viruses or malwares, it is very difficult to either to rectify the problem or to retrieve back the data as they are received from various types of sources. So, before storing the data and files in Big Data systems if we ensure that they are free from viruses and other malicious codes, it is a very good protection mechanism. This paper provides a methodology to identify viruses and other malicious codes in a Big Data system at an earlier stage.

Keywords: Virus, Worm Malware, Signature, Detection, Mechanism.

1 Introduction

Most commercial antivirus software products currently available use signature-based virus identification methods. However, the signature-based detection algorithms simply use the byte signatures of known viruses saved in memory to generate detection models. In general, detection methods based on byte signatures use a large collection of regular expressions or simple signature-string-matching engines to scan files. In addition to antivirus software products, packet processing applications that inspect packets at a level beyond protocol headers need to analyze the contents for some known signatures. For instance, network security applications must neglect packets containing certain harmful internet viruses and worms carried in packet payloads.

A computer virus is a computer program that can copy itself and infect another computer without the permission or knowledge of its user [2]. The computer viruses have the capability of spreading from machine to machine and is typically done without the user's knowledge or permission. Viruses add their code to other systems in such a way that whenever the infected part of the system executes, the viral code is also executed and the virus spreads further. An important primary characteristics of computer viruses is their ability of either to reproduce themselves or to produce an altered version of themselves.

A new zero day virus will not be detected by traditional detection systems unless this new virus is received by the antivirus companies and the virus signature is stored in their own databases. Signature-based detection systems need databases in order to store the signatures. As the number of viruses increases every day, huge databases are needed to store all their signatures, so that more storage space will be needed in the near future. The huge files in the databases will also affect the speed of searching for signatures, and, thus, affect the performance of the system. These disadvantages mean that the signature-based detection techniques will soon be inadequate to protect computer systems. Behavior-based virus detection systems have been developed recently. They do not rely on a database of signatures, but instead concentrate on the behavior of the system. They have come to light in order to overcome the problems associated with traditional signature-based detection. The principle behind this approach is first to observe the normal behavior of the system, after which any deviation from it will be classified as an intrusion. Next step in

this approach is to predefine virus behavior, so that any process which resembles virus activity can be identified as a potential virus. However, there are also some difficulties associated with behavior-based detection.

The computer virus writers use many strategies to evade detection like space filling, compression, encryption and other code transformation techniques. However; all the existing methods used by the available antivirus softwares are not adequate as new viruses are emerging rapidly [4].

One of the objective of this research is to safeguard the Big Data system from virus like threats. The new methodology used here provides a mechanism to identify virus like threats in Big Data systems at an earlier stage. This research provides a way to store only those files unaffected by virus and other threats in a Big Data system.

2 Literature Survey

Adrian Lane (2012) concluded that Big Data clusters are prone to threats which are similar to web applications and traditional data warehouses. He also described about the most common parallel processing and redundant storage issues in Big Data. He further questioned the security of the data and files in a Big Data system and concluded that the vulnerabilities not only confined to Big Data clusters but they may reside outside the clusters also. So, it is essential to improve security in Big Data systems against various vulnerabilities in the system [1].

Alvaro A. Cárdenas et al. (2013) investigated security problems from first generation 'Intrusion detection systems' to third generation 'Big Data analytics'. Their focus was on Big Data security and the use of cluster Infrastructures that made it more reliable and available still there remains a scope of improvement. Big data is changing the landscape of security technologies for network monitoring, SIEM (Security Information and Event Management) and forensics. However, in the eternal arms race of attack and defense, Big Data is not a panacea, and security researchers must keep exploring novel ways to contain sophisticated attackers. Big Data can also create a world where maintaining control over the revelation of our personal information is constantly challenged [2].

Eweka Raphael Osawaru et al. (2014) highlighted that the top security and privacy problems that need to be addressed to make Big Data processing and computing infrastructure more secure. Big Data security challenges are magnified by its own attributes like variety, volume, velocity, etc. Big Data has now become a very useful technology in the world where critical decisions made by Government and Organizations now depend on thorough analysis of both streaming and static large data sets. Its ability to produce historical and predictive results is overwhelming. However, all its benefits can be dwarf by its challenges, foremost of which is security. They carefully highlighted various security issues Big Data analytics faces so far and encourage further collaborative research for mitigating both security and privacy challenges relating to Big Data. As computing environments become cheaper, application environments become networked, system and analytics environments become shared over the cloud, security, access control, compression, encryption and compliance introduce challenges that must be addressed in a systematic way [3].

Venkata Narasimha Inukollu et al. (2014), concluded that data security not only involves the encryption of the data, but also ensures that appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms also have to be secure. The Big Data issues are most acutely felt in certain industries, such as telecoms, web marketing and advertising, retail and financial services, and certain government activities. The data explosion is going to make life difficult in many industries, and the companies will gain considerable advantage which is capable to adapt well and gain the ability to analyze such data explosions over those other companies. Finally, data mining techniques can be used in the malware detection in clouds [4].

Gautam Siwach et al. (2014) proposed an approach for identifying the encoding techniques in order to perform an expedited search over encrypted text for ensuring the security enhancements in Big Data. Data protection and security is of high importance and almost everyone's concern that is due to the sensitivity of data ranging from personal, financial details to the data containing national security. They also concluded that the data needs security at the database level [5].

Duygu Sinanc Terzi et al. (2015) examined the studies on Big Data security and privacy. They also presented the problems and solutions for security, privacy and forensic issues on Big Data and cloud computing. Cloud computing systems also support Big Data structures and solutions. In order to understand research projects better, Big Data sets were evaluated based on security. Here they presented security, privacy and forensic issues on these technological developments, and emphasized critical points on these issues. The results have shown that the cloud which supports Big Data systems is exposed to new threats [6].

Jagruti Parmar et al. (2016) reviewed the security issues and various approaches to overcome them in Big Data system. With the arrival of improved technologies, there has been exploding increase in the generation of data. When this large amount of data travels through the internet then there is the problem of securing, managing, storing and analyzing the data. The major issue then arises is the privacy and security of Big Data. Big data privacy, safety and security are the biggest issues to be discussed more in the future [7].

Mani Sarma Vittapu (2016) Proposed to design generic systems that can provide near real-time analytic services for many applications, such as spam detection, game log analysis, social community mining, etc. When data size keeps on increasing, more complex user requirements need to be handled. The emergence of new hardware violates the old design and the old system becomes too complicated for maintenance. According to him no solution can address all Big Data problems [8].

3. Objective:

A Big Data system is used to store variety of data of large volumes from heterogeneous sources. If virus affected files are stored knowingly or unknowingly in any part of the Big Data system it generates huge problems by affecting all other files those are of huge volume. The objective of this paper is to identify virus like threats in Big Data system at an earlier stage. If we restrict the system at the beginning itself to filter these virus affected files, we can protect the Big Data system from these threats easily.

4. Existing Methodology

A basic requirement of Big Data storage system is to protect the data and files from intruders, malicious codes and other attacks. There are some existing mechanisms like Attribute Based Encryption, Identity Based Encryption, Homomorphic Encryption, Storage Path Encryption, etc., to provide security in these systems.

In an Attribute Based Encryption system access polices are defined by data owner and data are encrypted under those policies. The data can only be decrypted by the users whose attributes satisfy the access policies defined by the data owner.

Identity Based Encryption is an alternative to Public Key Encryption (PKE) which is proposed to simplify key management in a certificate-based Public Key Infrastructure (PKI) by using human identities like email address or IP address as public keys.

Another approach to support the updating of ciphertext at the receiver end is to delegate this task to a trusted third party.

Proxy-Re Encryption (PRE) is another approach that was proposed to handle the problem of data sharing between different recipients. Here a semi trusted third party transforms a ciphertext intended for one user into that intended for another user without leaking any knowledge about the message or the decryption keys. The workload of data owner is now transferred to the proxy and the proxy does not have to be online all the time.

Usage of Hybrid Clouds is another approach in which hybrid clouds are deployed. According to the national institute of standards and technology (NIST), the cloud can be deployed as private cloud or public cloud or hybrid cloud. Private clouds are inherently trustworthy and secure but there are some limitations which hamper the private clouds for the processing and storage of Big Data.

The first limitation is scalability. Building a highly scalable private cloud requires a large capital investment. It becomes very difficult to accurately plan private cloud capacity when the volume, velocity, and variety of the data are constantly changing. The second limitation is unavailability of analytical models and software frameworks required to manage heterogeneous data. The third limitation is on data sharing. Sometimes, data sharing should be available among authorized collaborators who do not have access or reside outside of private cloud. However, due to security concerns, this is not always possible. On the other hand, public cloud support scalability and easy sharing of data. However public clouds are more prone to security and privacy attacks.

5. Proposed Work:

When dealing with Big Data one may often need to change data access policies as the data owner may have to share it with different organizations. The attribute based access control schemes do not consider policy updating. The policy updating is a very challenging task in attribute based access control systems.

Encryption schemes like Identity Based Encryption(IBE) and Attribute Based Encryption(ABE) did not support updating of ciphertext at the receiver end as the size of the data are large in Big Data System, the decryption and re-encryption can be very time consuming and costly because of computation overhead. Moreover, in this mode, data owner has to be online all the time. The trusted third party approach has also drawbacks as the scheme depends on the complete trust of the third party.

In public key cryptography the major problem is the key lengths. The lengths of the keys which are used in public key encryption are very large. This results in the low speed of transmission of the data. It is difficult to them to identify threats due to viruses at an initial stage and there are chances for the virus threats to affect the keys.

The present virus detection methods use large databases and different techniques to detect viruses. The important drawback of these methods is that they take more time and more memory space. In present day to day scenario Big Data systems are efficient systems to store and retrieve vast data from heterogeneous data sources. Our proposed methodology provides a mechanism to identify virus affected files using code values in Big Data systems.

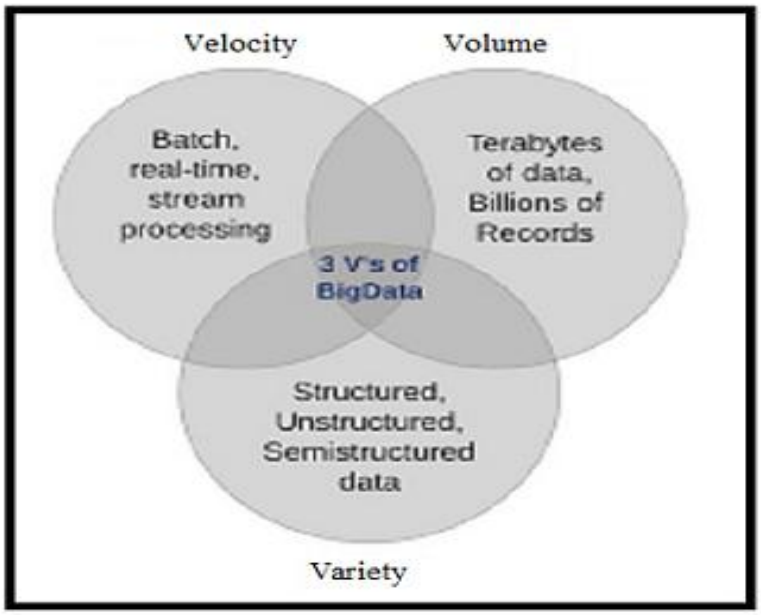

Figure 1: schematic representation of the 3V‘s of Big Data

6Design and Implementation

Here we proposed a methodology in which the files to be stored in Big Data system are initially checked for virus like threats. This can be done by checking the code values which depends on the basic parameters of file system. The algorithm used in the proposed methodology is given as follows:

Step 1: Initiate
Step 2: Input Source file
Step 3: Calculate code values SF1 and DF1 for the file at source and destination systems.
Step 4: Compare both the code values SF1 and DF1.
Step 5: If the code values of both files match then send it to Big Data system.
Step 6: Otherwise repair/delete the file at destination.
Step 7: Initiate virus scanning and cleaning activities at source and destination.
Step 8: Request to resend the file.
Step 9: End.

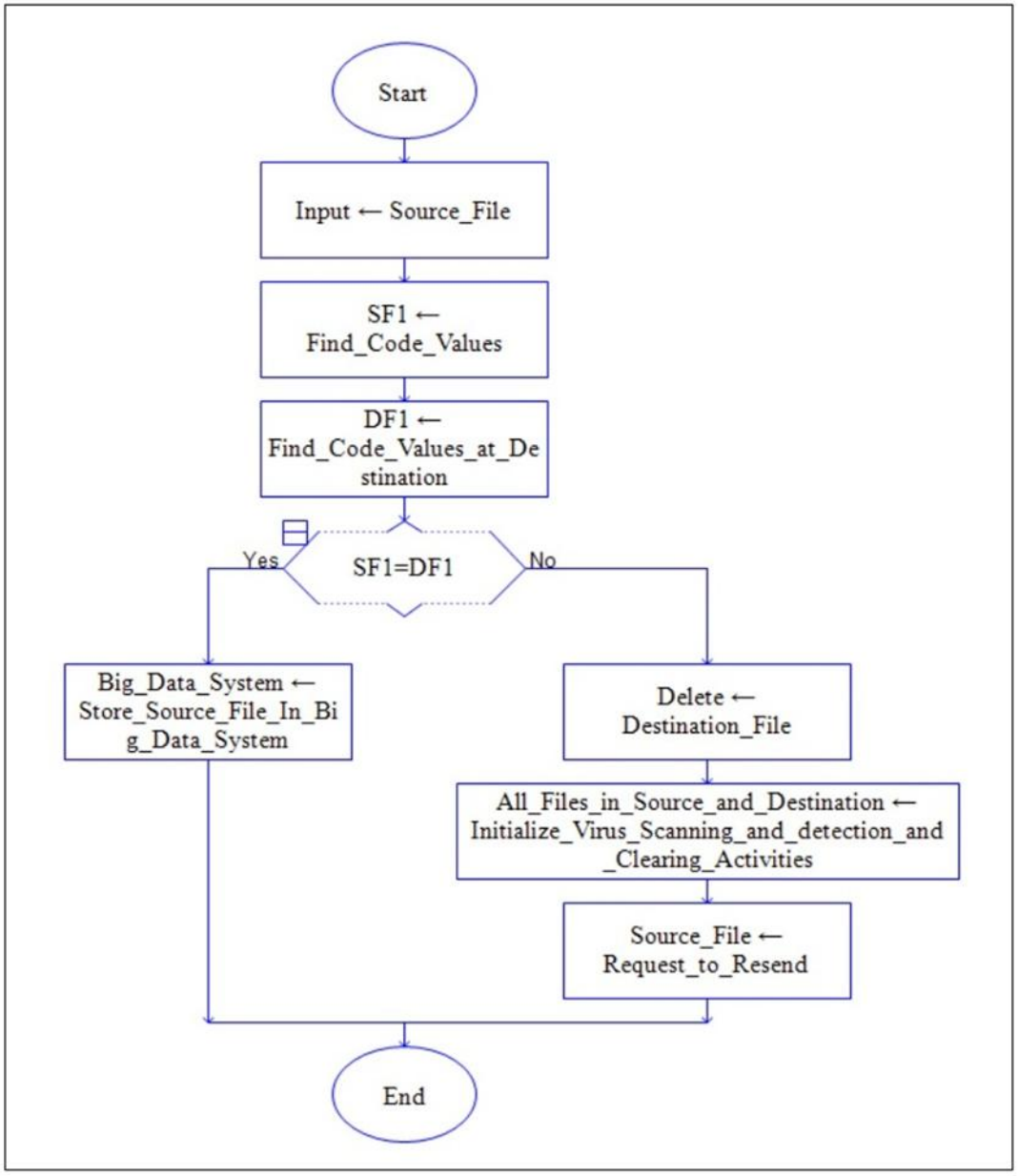

Figure 2: Flow chart of the proposed Virus Threat Identification methodology in Big Data system.

The flowchart given above shows the methodology used in the design of the proposed system.

7 Experimentation and Results

The code values obtained in our experiments on test files and the results are tabulated and given in the following Table 1

7Experimentation and Results:

The sample experiments using the proposed method were carried out and the results were shown in the following Table 1.

S.No	File name	Code values in		Conclusion
		Source system	Destination system	
1	V_Basic.txt	A459E548E2F0BA04 468BBF4AB6B5394C	262813B73BE46E2692 CFD1FAE0D3E384	**Threats Identified. Virus alert and Scanning initiated**
2	SNR_photo1. jpg	957D1FBA9A6960E 536DFD584F802AB17	957D1FBA9A6960E53 6DFD584F802AB17	No Threat
3	scratch_sampl e.sb	74402B64C1B25F63 32B851D43CFE3E92	95626538A23D1689AF B6CC9A654D9F76	**Threats Identified. Virus alert and Scanning initiated**
4	ijcsit2014050 3265.pdf	2BEBE7B5512ED3E 60477A6EE1F990423	2BEBE7B5512ED3E60 477A6EE1F990423	No Threat
5	Database21.a ccdb	B1BBB2126CA08E9 96E5FF9D10112E3FE	380A27D50111B9F7E 92388411D24F78A	**Threats Identified. Virus alert and Scanning initiated**

Table 1: Code values for the sample files in Big Data system.

In the Table 1 the files V_Basic.txt and scratch_sample.sb were affected by virus like threats and the remaining files are not affected by such threats. Thus using the above methodology virus like threats were identified.

8 Conclusion

Today the existing Big Data Systems are facing lot of threats due to viruses and other malicious codes. The existing virus scanners cannot provide guarantee to detect all viruses. The important feature of the proposed method is that it uses code values to distinguish normal files from virus affected files at an earlier stage. Virus alert may be sent and scanning may be initiated immediately to protect the entire Big Data System. After this process genuine files unaffected by virus and other malwares can be stored in Big Data system. So, the proposed methodology provides a way to store virus free files to Big Data systems and avoids vast problems due to various threats.

Thus, in our experiments the results showed that the files affected by virus and other malware threats at different locations in a Big Data System were identified successfully.

9 Future Work

The above proposed method is an efficient method to store files in a Big Data system. This method may be considered for further research to get improvements in storing virus free files in Big Data systems.

10 References

[1] Adrian Lane, (2012) 'Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments' https://securosis.com/assets/ library/reports/ Securing Big Data_ FINAL.pdf
International Journal of Computing & Information Sciences Vol. 12, No. 2, December 2016

[2] Alvaro A. Cárdenas,Pratyusa K. Manadhata and Sreeranga P. Rajan, (2013), 'Big Data Analytics for Security', Copublished by the IEEE Computer and Reliability Societies.

[3] Eweka Raphael Osawaru and Riyaz Ahamed A. H., (2014), 'A Highlight of Security Challenges in Big Data', International Journal of Information Systems and Engineering (online), Volume 2, Issue 1,ISSN: 2289-2265 Page 1.

[4] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, (2014), 'Security Issues Associated with Big Data in Cloud Computing', International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3.

[5] Gautam Siwach and Amir Esmailpour, (2014), 'Encrypted Search & Cluster Formation in Big Data', ASEE 2014 Zone I Conference, April 3-5, 2014, University of Bridgeport, Bridgpeort, CT, USA.

[6] Duygu SINANC TERZI, Anıl AYAYDIN and Seref SAGIROGLU, (2016), 'Security, Privacy and Forensics Issues on Big Data and Cloud Computing', https://www.researchgate.net/ publication/ 308778648

[7] Jagruti Parmar and Jalpa Patel, (2016), 'A Review of Big Data Security Issues and Approaches', International Journal for Scientific Research & Development| Vol. 4, Issue 07, 2016 | ISSN (online): 2321-0613.

[8] Mani Sarma Vittapu, (2016), 'Design Approach to Big Data Systems in Developing and Maintaining the Information Security Systems', IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. III www.iosrjournals.org