isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 2 ,Apr,2023 ,ISSN_NO: 2321-3337

EFFECTIVE BIG DATA ANALYSIS TOWARDS ATTACK DETECTIONIN CLOUD ENVIRONMENT FOR SECURED DATA ACCESBILITIES

BHAVYASHREE.G[1], DEEPIKA .C[2], PAVITHRA. L[3], EDITH ESTHER. E[4]

UG Scholar[1-3] Department of Computer Science, GRT Institute of Engineering and Technology, Tiruttani, India.

Assistant Professor[4] Department of Computer Science, GRT Institute of Engineering and Technology, Tiruttani, India.

shreebhavya287@gmail.com, deepikachandrasekar32@gmail.com, pavithraloganathan15@gmail.com ,

edithesther@gmail.com

Abstract - we are implementing a Big Data based centralized log analysis system to identify the network traffic occurred by attackers through DDOS, SQL Injection and Brute Force attack. The log file is automatically transmitted to the centralized cloud server and big data is initiated for analysis process. To implement DDOS attack is the continuous request from the same IP to avoid the Cloud server to function normally. Brute Force Attack is providing the fake / wrong Passwords for accessing the cloud server. SQL Injection is given by the SQL itself to the admin to access the User Accounts this, Hacker logins into the server by providing wrong Password or 1 = 1 in the password field. Generally, this will allow the hacker to get into the user's Account. So, we are identifying all the three Attacks through our Application by generation of Log file which is uploaded to Big data for Attack detection.

1.INTRODUCTION

Virtualized infrastructure consists of virtual machines (VMs) that rely upon the software-defined multi-instance resources of the hosting hardware. The virtual regulates, and manages the software-defined multi-instance architecture. The ability to pool different computing resources as well as enable on-demand resource scaling has led to the widespread deployment of virtualized infrastructures as an important provisioning to cloud computing services.

This has made virtualized infrastructures become an attractive target for cyber attackers to launch attacks for illegal access. Exploiting the software vulnerabilities within the hypervisor source code, sophisticated attacks such as Virtualized Environment Neglected Operations Manipulation (VENOM) have been performed which allow an attacker to break out of a guest VM and access the underlying hypervisor. In addition, attacks such as Heartbleed and Shellshock which exploit the vulnerabilities within the operating system can also be used against the virtualized infrastructure to obtain login details of the guest VMs and perform attacks ranging from privilege escalation to Distributed Denial of Service (DDoS).

Existing security approaches to protecting virtualized infrastructures generally include two types, namely malware detection and security analytics. Malware detection usually involves two steps, first, monitoring hooks are placed at different points within the virtualized infrastructure, then a regularly-updated attack signature database is used to determine attack presence. While this allows for a real-time detection of attacks, the use of a dedicated signature database makes it vulnerable to zero-day attacks for which it has no attack signatures.

2. BACKGROUND

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. The challenges include analysis, capture, duration, search, sharing, storage, transfer, visualization, and privacy violations.

2.1. OVERVIEW OF BIG DATA

The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on. So we can implement big data in our project because every employ has instructed information so we can make analysis on this data.

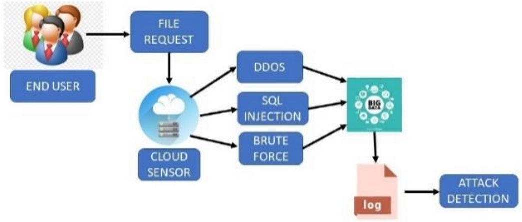


FIG 1: ATTACK DETECTION RELATED TO BIGDATA

SOME CHARACTERISTICS OF BIG DATA

Big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process, and analyse through traditional database technologies. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insights. The term "big data" is relatively new in IT and business. However, several researchers and practitioners have utilized the term in previous literature. For instance, referred to big data as a large volume of scientific data for visualization. Several definitions of big data currently exist." Meanwhile and defined big data as characterized by three Vs: volume, variety, and velocity. The terms volume, variety, and velocity were originally introduced by Gartner to describe the elements of big data challenges. IDC also defined big data technologies as a new technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis." specified that big data is not only characterized by the three Vs mentioned above but may also extend to four Vs, namely, volume, variety, velocity, and value This 4V definition is widely recognized because it highlights the meaning and necessity of big data.

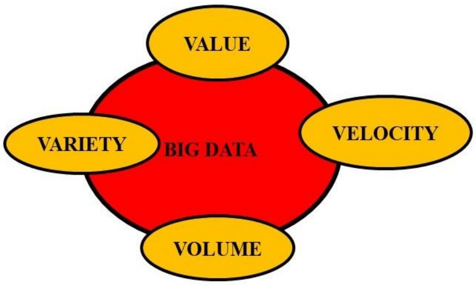


FIG 2.1. FOUR V'S OF BIG DATA

Volume - Refers to the amount of all types of data generated from different sources and continue to expand. The benefit of gathering large amounts of data includes the creation of hidden information and patterns through data analysis Collecting longitudinal data requires considerable effort and underlying investments. Nevertheless, such mobile data challenge produced an interesting result similar to that in the examination of the predictability of human behaviour patterns or means to share data based on human mobility and visualization techniques for complex data.

Variety-Refers to the different types of data collected via sensors, smart phones, or social networks. Such data types include video, image, text, audio, and data logs, in either structured or unstructured format. Most of the data generated from mobile applications are in unstructured format. For example, text messages, online games, blogs, and social media generate different types of unstructured data through mobile devices and sensors. Internet users also generate an extremely diverse set of structured and unstructured data.

Velocity -Refers to the speed of data transfer. The contents of data constantly change because of the absorption of complementary data collections, introduction of previously archived data or legacy collections, and streamed data arriving from multiple sources

Value- is the most important aspect of big data; it refers to the process of discovering huge hidden values from large datasets with various types and rapid generation.

3. RELATED WORK

Big data are classified into different categories to better understand their characteristics shows the numerous categories of big data. The classification is important because of large-scale data in the cloud. The classification is based on five aspects: (i) data sources, (ii) content format, (iii) data stores, (iv) data staging, and (v) data processing.

Data sources include internet data, sensing and all stores of transnational information, ranges from unstructured to highly structured are stored in various formats. Most popular is the relational database that come in many varieties. As the result of the wide variety of data sources, the captured data differ in size with respect to redundancy, consistency and noise, etc.

Big Data Storage System

The rapid growth of data has restricted the capability of existing storage technologies to store and manage data. Over the past few years, traditional storage systems have been utilized to store data through structured RDBMS. However, almost storage systems have limitations and are inapplicable to the storage and management of big data. A storage architecture that can be accessed in a highly efficient manner while achieving availability and reliability is required to store and manage large datasets.

Several storage technologies have been developed to meet the demands of massive data. Existing technologies can be classified as direct attached storage (DAS), network attached storage (NAS), and storage area network (SAN). In DAS, various hard disk drives (HDDs) are directly connected to the servers. Each HDD receives a certain amount of input/output (I/O) resource, which is managed by individual applications. Therefore, DAS is suitable only for servers that are interconnected on a small scale. Given the aforesaid low scalability, storage capacity is increased but expandability and upgradeability are limited significantly. NAS is a storage device that supports a network. NAS is connected directly to a network through a switch or hub via TCP/IP protocols. In NAS, data are transferred as files. Given that the NAS server can indirectly access a storage device through networks, the I/O burden on a NAS server is significantly lighter than that on a DAS server. NAS can orient networks, particularly scalable and bandwidth-intensive networks. Such networks include high-speed networks of optical-fiber connections.

The SAN system of data storage is independent with respect to storage on the local area network (LAN). Multipath data switching is conducted among internal nodes to maximize data management and sharing. The organizational systems of data storages (DAS, NAS, and SAN) can be divided into three parts: (i) disc array, where the foundation of a storage system provides the fundamental guarantee, (ii) connection and network subsystems, which connect one or more disc arrays and servers, and (iii) storage management software, which oversees data sharing, storage management, and disaster recovery tasks for multiple servers.

Hadoop is an open-source Apache Software Foundation project written in Java that enables the distributed processing of large datasets across clusters of commodity. Hadoop has two primary components, namely, HDFS and Map Reduce programming framework. The most significant feature of Hadoop is that HDFS and Map Reduce are closely related to each other; each are co-deployed such that a single cluster is produced. Therefore, the storage system is not physically separated from the processing system.

HDFS is a distributed file system designed to run on top of the local file systems of the cluster nodes and store extremely large files suitable for streaming data access. HDFS is highly fault tolerant and can scale up from a single server to thousands of machines, each offering local computation and storage. HDFS consists of two types of nodes, namely, a name node called “master” and several data nodes called “slaves.” HDFS can also include secondary name nodes. The name node manages the hierarchy of file systems and director namespace (i.e., metadata). File systems are presented in a form of name node that registers attributes, such as access time, modification, permission, and disk space quotas.

Map Reduce accelerates the processing of large amounts of data in a cloud; thus, Map Reduce, is the preferred computation model of cloud providers. Map Reduce is a popular cloud computing framework that robotically performs scalable distributed applications and provides an interface that allows for parallelization and distributed computing in a cluster of servers. The approach is to apply scientific computing problems to the Map Reduce framework where scientists can efficiently utilize existing resources in the cloud to solve computationally large-scale scientific data.

Currently, many alternative solutions are available to deploy Map Reduce in cloud environments; these solutions include using cloud Map Reduce runtimes that maximize cloud infrastructure services, using Map Reduce as a service, or setting up one’s own Map Reduce cluster in cloud instances. Several strategies have been proposed to improve the performance of big data processing. Moreover, effort has been exerted to develop SQL interfaces in the Map Reduce framework to assist programmers who prefer to use SQL as a high-level language to express their task while leaving all of the execution optimization details to the backend.

4. PROPOSED SYSTEM

We are implementing a system to identify the network traffic occurred by attackers and identify the attackers who is attacking the server. We are implementing a Big Data based centralized log analysis system to identify the network traffic occurred by attackers through DDOS, SQL Injection and Brute Force attack. The log file is automatically transmitted to the centralized cloud server and big data is initiated for analysis process.

4.1 Cloud Establishment

Data owner will upload their data to the cloud server and request for a particular file is send to cloud server. Both the upload and the file request are handled by the main Cloud Server. During the file request is processed main server will communicate with the data owner and the files are retrieved only after the approval given the data owner.

In our Project, we are using Drop box as the cloud server. We are integrating the Master key and secret key in our Java front end coding to connect directly to the cloud server. This Process of cloud connectivity is called as Infra Structure As A Service (IAAS). In IAAS data storage, access, File management and Infrastructural support will be provided to the user for better resource management.

isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 2 ,Apr,2023 ,ISSN_NO: 2321-3337

4.2 User Cloud Access

In this Module User interface is created so that the data owner will upload the data to the server. The main objective of this module is to store and share the data by uploading the file to the remote machine and maintain the file on cloud using User Interface page.

User is allowed to register into our Application to access the cloud. User will be allowed to access the files and all the user access information are stored and monitored by the continuous recordings into the Log File. This Log file is used for the complete activity monitoring using Big data.

4.3 Log File Generation

In this module, we have to create the log file for the user's searching information. The aim to implement this module is to generate the log file based on user searching, uploading and downloading information. Whenever user access the cloud server our system will automatically create the log for each and every accessing information. Those log files will create and stored on local memory as a json or text file.

This Log file generation used for the User Analysis using Big Data – HDFS. In general this Log file records storage and analysis plays a Vital Role towards security. Because in general, Cloud can be easily accessed by anywhere the world. So there are more possibilities for the attacks. Using this Log file, our System will detect the Attacks and protect the Cloud Data.

4.4 Log File Analysis Using Big Data

In this we use big data to analyse the log file. Using we generate a log file in the form of text or json file. Because, big data will support unstructured database in huge level. After the user's accessing information is gathered those data will generate as form of input. When big data executes its job, every slave will assign for a job and finally it will show the result as a output in the form of json or text file. It will be stored on local memory and it will automatically upload on cloud.

Big data Role will be really useful in the case of adding multiple Attacks and more data Analysis. In General, for the small volume of data, Big data is not needed, but Big data is integrated for the Further addition of Data and more number of attacks. We are using Hadoop Distributed File System, in this Project. We deploy Data Node, Name Node, Resource Manager & Node Manager in this Project.

4.5 Attack Analysis And Categorization

This module will play a vital role in this project, because the aim of the project implemented in this module. We implement the system to identify attacks on cloud. We maintain a set of attacks dataset and also we have a log file authorized user's accessing information. If unauthorized user injects any data on user's cloud server or try to inject any attack file on cloud server. We will compare those accessing information with dataset and categorize the attack file and user's file in different folder. Using big data we categorize the file.

We detect Three different Attacks in this Project, namely

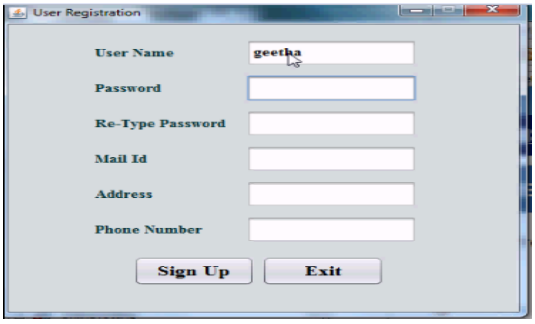
DDOS (Distributed Denial of Service) Attacks, which represents, Multiple File request from the same within the short span of time, which in turn disturb the Cloud Server's Response to the legitimate users.

SQL Injection Attacks is nothing but data accessibility rights to the Legitimate Users- Admins. The main concept of SQL Injection is legalized data Access, but eventually

Attackers used this concept to Hack the data.

In Brute Force Attack, Hacker will try with multiple passwords to match and get into the system

5. Experimental Output



User Name geetha

Fig.5.1. User Registration

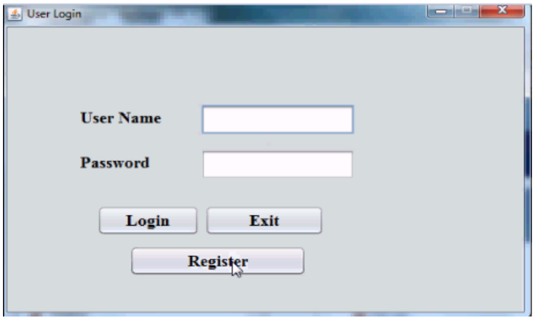
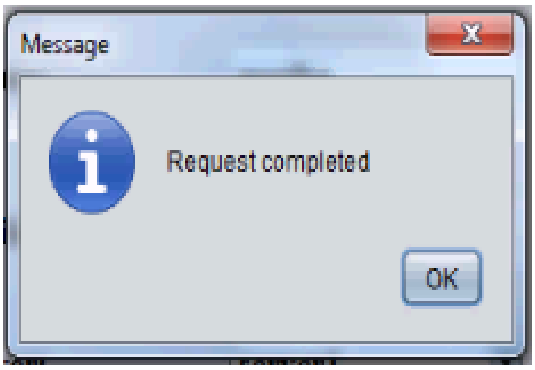


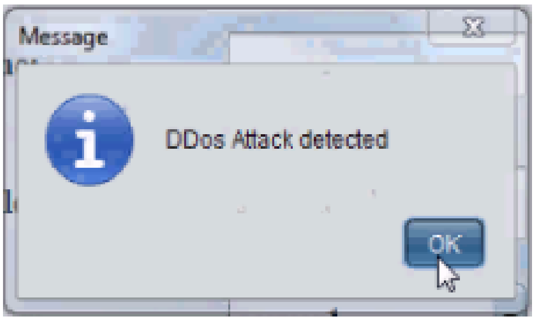
Fig.5.2. User Login



Message

Request completed

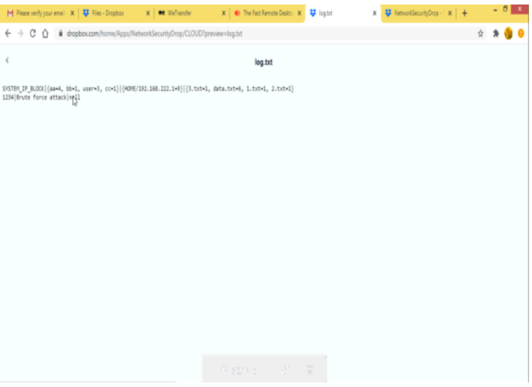
Fig.5.3. Request Message



Message

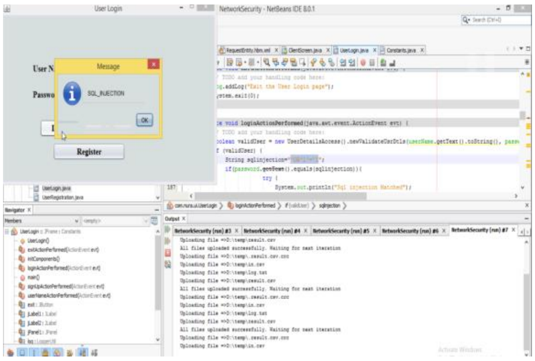
DDos Attack detected

Fig.5.4. Attack Detection



```
SYSTEM_IP_BLOCK|{aa=4, bb=1, user=3, cc=1}|{HOME/192.168.222.1=9}|{3.txt=1, data.txt=6, 1.txt=1, 2.txt=1}
1234|Brute force attack|null
```

Fig.5.5. Log Text

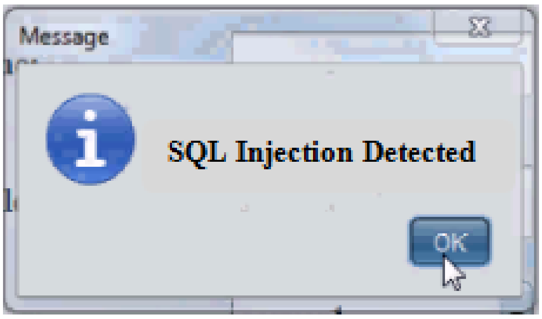


```
// TODO add your handling code here:
log.addLog("Exit the User Login page");
System.exit(0);

private void loginActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
    boolean validUser = new UserDetailsAccess().newValidateUsrDtls(userName.getText().toString(), pass
    if (validUser) {
        String sqlinjection="'OR'1'='1";
        if(password.getText().equals(sqlinjection)){
            try {
                System.out.println("Sql injection Matched");
```

```
Uploading file =>D:\temp\result.csv
All files uploaded successfully. Waiting for next iteration
Uploading file =>D:\temp\.result.csv.crc
Uploading file =>D:\temp\in.csv
Uploading file =>D:\temp\log.txt
Uploading file =>D:\temp\result.csv
All files uploaded successfully. Waiting for next iteration
Uploading file =>D:\temp\.result.csv.crc
Uploading file =>D:\temp\in.csv
Uploading file =>D:\temp\log.txt
Uploading file =>D:\temp\result.csv
All files uploaded successfully. Waiting for next iteration
Uploading file =>D:\temp\.result.csv.crc
Uploading file =>D:\temp\in.csv
```

Fig.5.6. Network Security



Message

SQL Injection Detected

Fig.5.7. SQL Injection Detection

6. CONCLUSION

The aim of this Project, is to identify the Attackers in the cloud Environment, as Cloud is more susceptible for Attacks. Our System is well designed to identify three different types of Attacks namely, DDOS, SQL Injection & Brute Force Attack. We have also integrated Big Data for further addition of more Attacks in Future. Once the Attacks are identified using big data, our system will IP Address and finally ensures Security in the Cloud environment.

7. References

[1] D. Fisher, "'venom' flaw in virtualization software could lead to VM escapes, data theft," 2015. [Online]. Available: https://threatpost.com/venom-flaw-in-virtualization-software-could-lead-tome- escapes-data-theft/112772/, Accessed on: May 20, 2015.

[2] Z. Durum Eric, et al., "The matter of Heart bleed," in Proc. Conf. Internet Meas. Conf., 2014, pp. 475–488.

[3] K. Cabaj, K. Grochowski, and P. Gawronski, "Practical problems of internet threats analyses," in Theory and Engineering of Complex Systems and Dependability. Berlin, Germany: Springer, 2015, pp. 87–96.

[4] J. Overhead, E. Cooke, and F. Johann Ian, "Cloud AV: N-version antivirus in the network cloud," in Proc. USENIX Secure. Symp., 2008, pp. 91–106.

[5] X. Wang, Y. Yang, and Y. Zeng, "Accurate mobile malware detection and classification in the cloud," Springer Plus, vol. 4, no. 1, pp. 1–23, 2015.

[6] P. K. Chouhan, M. Hagan, G. McWilliams, and S. Sezer, "Network based malware detection within virtualised environments," in Proc. Eur. Conf. Parallel Process., 2014, pp. 335–346.