iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 Issue: IV Month of publication: April 2017

DOI: http://doi.org/10.22214/ijraset.2017.4065

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# Intrusion Detection Techniques in Big Data: A Review

Neenu Daniel[1], Ritty Jacob[2]
[1,2]*Department of CSE, VJCET, Vazhakulam,Kerala*

***Abstract**: With the increasing use and demand of Big Data, Security has become the major concern. With the growth of cyber-attacks, information safety has become an important issue all over the world. Prevention alone is not the solution for this; hence Intrusion Detection Systems have come into focus. . Intrusion refers to the hacking of the system or a network which poses a security risk to the private information of the organization. Intrusion detection systems (IDSs) are an essential element for network security infrastructure and play a very important role in detecting large number of attacks..This paper introduces a detailed analysis of the network security problems and also represents a review of the current research in Big Data intrusion detection systems. In this paper an attempt has been made to identify recent approaches used for intrusion detection in Big Data.*

**Keywords***: BIG DATA, IDS, HIDS, NIDS*

## I. INTRODUCTION

Big Data is a data analysis methodology enabled by a new generation of technologies and architecture which support high-velocity data capture, storage, and analysis (Villars, Olofson, & Eastwood, 2011). Big Data requires huge amounts of storage space. As the amount of data being collected continues to grow, more and more companies are building big data repositories to store, aggregate and extract meaning from their data. Policies related to privacy, security, intellectual property, and even liability will need to be addressed in a big data world. With Big Data databases, enterprises can save money, grow revenue, and achieve many other business objectives by building new applications, improving the effectiveness and lowering the cost of existing applications, realizing new sources of competitive advantage. The continuous collection of traffic data by the network leads to Big Data problems that are caused by the volume, variety and velocity properties of Big Data. The use of large scale cloud infrastructures, with a diversity of software platforms, spread across large networks of computers, also increases the attack surface of the entire system

The major contribution of this paper includes various security threats, intrusion detection approaches and techniques in Big Data environment. Section 2 gives the detailed information about the Big Data and Security. Intrusion detection basics are discussed in section 3. Section 4 discusses about the various intrusion detection Approaches. Recent Intrusion Detection techniques in Big Data are given in Section 5 followed by the conclusion section.

## II. SECURITY IN BIG DATA

### *A. Big Data*

Big data is a term that describes the large volume of data – both structured and unstructured – that overwhelm a business on a day to day basis[1]. But it’s not the amount of data that’s important. It’s what organizations do with the data that matters. This concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs:Volume. Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would’ve been a problem – but new technologies (such as Hadoop) have eased the burden. Velocity. Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

### *B. Big Data Security Challenges*

Although our data capacity is growing exponentially, we have imperfect solutions for the many security issues .The challenges associated with big data privacy issues which can be divided into four groups[5]:

1) *Infrastructure Security:* Secure computations in distributed programming frameworks as well as in non relational data stores. Distributed programming frameworks process big data with parallel computation and storage techniques. In such frameworks, unauthenticated or modified mappers which divide huge tasks into smaller sub-tasks so that the tasks can be aggregated to

create a final output — can compromise data. Faulty or modified worker nodes which take inputs from the mapper to execute the tasks — can compromise data by tapping data communication between the mapper and other worker nodes. Most cloud-based data frameworks use the NoSQL database. The NoSQL database is beneficial for handling huge, unstructured data sets but from a security perspective, it is poorly designed. NoSQL was originally designed with almost no security considerations in mind. One of the biggest weaknesses of NoSQL is transactional integrity. It has poor authentication mechanisms, which makes it vulnerable to man-in-the-middle or replay attacks. To make things worse, NoSQL does not support third-party module integration to strengthen authentication mechanisms. Since authentication mechanisms are rather no strict, data is also exposed to insider attacks. Attacks could go unnoticed and untracked because of poor logging and log analysis mechanisms

2) *Data Privacy:* Secure the data itself using a privacy-preserving approach for data mining and analytics. Also, protect sensitive data through the use of cryptographically enforced datacentric security and granular access control. The amount of information collected on each individual can be processed to provide a surprisingly complete picture. As a result, organizations that own data are legally responsible for the security and the usage policies they apply to their data. Attempts to anonymous specific data are not successful in protecting privacy because there is so much available that some data can be used as a correlation for identification purposes. Users' data are also constantly in transit, being accessed by inside users and outside contractors, government agencies, and business partners sharing data for research.
3) *Data Management*: Manage the enormous volume of data using scalable, distributed solutions to secure data stores and enable efficient audits and data provenance. There are specific vulnerabilities associated with big data storage: Confidentiality and integrity, data provenance, and consistency.
4) *Integrity/ Reactive Security:* Use endpoint validation and filtering to check the integrity of streaming data, and real-time security monitoring and analytics to help prevent and address security problems. Because of the breadth of data sources, including endpoint collection devices, a major challenge facing big data schemes is whether the data is valid from the point of input. Given the size of the data pool, how can we validate the sources? How can we be sure that a source of input data is not malicious, or simply incorrect? In addition, how can we filter out malicious or unreliable data? Both data collection devices and programs are susceptible to attack. An infiltrator may spoof multiple IDs and feed fake data to the collection system.

### C. *Tools for Big Data Analysis*

Hadoop is a software framework for storing and processing Big Data and work under Big Data Analytics. It is an open source framework build on java platform and aimed at to improve the performance in terms of data processing on Big Data[3].

Features of Hadoop:

Hadoop has two core components: HDFS and Map Reduce. HDFS is used for storing huge data sets while Map Reduce is used for processing these huge data sets.Hadoop consists of multiple concepts like COMBINER, PARTITIONER HBASE, PIG, HIVE, SQOOP to perform the easy and fast processing of huge data sets.

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity. MapReduce is a parallel proramming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multiterabyte datasets), on large clusters (thousands of nodes) of commodityhardware in a reliable, faulttolerant manner. The MapReduce program runs on Hadoop which is an Apache opensource framework.

## III. INTRUSION DETECTION SYSTEM

In a network or a system any kind of unauthorized activities called intrusions. An intrusion detection system is a collection of tools, methods, and resources to help identify access and report intrusions. IDSs produce alerts for the administrators which are based on true positives or true alarms when actually intrusion takes place and false positive or false alarms in case of a wrong detection by the system. Mainly there are three types of IDS in cloud computing systems: Host based IDS, Network based IDS. Host based IDS(HIDS) monitors specific host machines, network-based IDS (NIDS) identifies intrusions on key network points

### A. *Host-based Intrusion Detection Systems*

HIDS analyses the traffic to and from the specific computer on which the intrusion detection software is installed. A host-based system also has the ability to monitor key system files and any attempt to overwrite these files.

### *B. Network -Based Intrusion Detection Systems*

A NIDS is often a standalone hardware appliance that includes network detection capabilities. It will usually consist of hardware sensors located at various points along the network. It may also consist of software that is installed on various computers connected along the network. The NIDS analyses data packets both inbound and outbound and offer real-time detection.

## IV. INTRUSION DETECTION APPROACHES

IDS can be classified into two detection approaches: misuse detection and anomaly detection. Misuse detection approach monitors network traffic or system activities for known misuse, most of the case using table of pattern called signatures. IDS will match the event with the signature to detect the event as misuse or not. Anomaly detection approach on the other hand, detects any intrusion based on its decision using some techniques including statistical and machine learning. The IDS will first learn about the normal behavior of the network or system and create a profile of it. If the is any event that did not match the profile is considered anomalous.

## V. INTRUSION DETECTION TECHNIQUES

Many researchers have suggested several IDS especially for the Big Data. Some of them will be reviewed in the following paragraph.

### *A. Hybrid Scheme based on Big Data Analytics IDS*

In this work Heterogeneous data from different sources has been collected from KDD Cup Dataset and segregated into learning phase and detection phase[2]. In the learning phase, known attacks will be identified. Similarly detection phase also will consider the same. The proposed system specifies a set of rules and high DoS, R2L, U2R, Probe. In this detection phase using enhanced C4.5 and enhanced Genetic Algorithm were used. These two techniques were storing the data in-terms of database independently. These two databases were hybrid together forming an integrated large database. The output of this large database will be applied as an input to Big-data using data streams one pass constraint technique using this technique data sets are classified. corresponding to this dataset we can find the relative type of attack which are DoS, Probe, R2L and U2R.

### *B. Clustering Based IDS*

Sonal Ashok Hajare proposed a Intrusion detection algorithm based on big data analysis[6]. In this paper Big data analysis framework use map reduced Intrusion detection system based on clustering algorithm. In this work data is collected from different sources and perform preprocessing on data with Fuzzy C Means clustering algorithm. Finally data analysis is done using support vector machine algorithm which is used for detection of network attacks. This technique is suitable to detect unknown attacks.

### *C. Snort and Hadoop based IDS*

In this work[9], the intrusion detection system Snort is used. The packets captured by Snort are analyzed by the Grid computing framework Hadoop, which is used for Big Data Analysis. For more user friendlier analysis a data warehouse system for Hadoop, Hive is also provided. For those ip addresses that generate large number of packets, Snort rules will be generated so that when the number of packets from a particular source exceeds a number, the node will generate alerts to other nodes since there is a possibility of attack. Snort rules generated after analysis very much efficient in detecting many attacks.

### *D. LDA based IDS*

This paper explores a hybrid approach of intrusion detection through knowledge discovery intrusion detection through knowledge discovery from big data using Latent Dirichlet Allocation. LDA is a powerful topic modeling technique developed in NaturalLanguageProcessing and Machine Learning. Signature-based technologies catch the known attacks effectively, but are not effective to the rapid growing new types of attacks; anomaly-based technologies detect “abnormal” behaviors, therefore this approach may catch the unknown attacks but typically has high false positive rate. In this paper the security incidents identified by LDA catch the latent semantics of a collection of the logged events, which gives a higher level description about what a user is doing or intends to do by that specific set (or sequence) of operations. This latent semantics is critical to intrusion detection, and can be identified by LDA modelling.

### *E. Hadoop and Naive Bayesian Classifier based IDS*

In this work[12] naive Bayesian based intrusion detection is proposed. The intrusion detection system that employs a Naïve Bayes algorithm that runs in a distributed manner. The classifier uses the Apache Hadoop and HStreamingAPIs to detect intrusions in real time. The network traffic data generated by noncommercial packet sniffers serve as input to classifier.  Its aim is to develop a distributed, scalable, fault tolerant and reliable system. The system is designed in such way that it can run effortlessly on a cluster comprised of obsolete hardware. The results denote that proposed system's performance is acceptable in real case scenarios.

### *F. Big Data Classification: Problems and Challenges in Network Intrusion*

Shan Sutharan [4] proposed a method on the specific problem of Big Data classification of network intrusion traffic. The first challenging problem rests on the current definition of Big Data; how to prove that the network traffic data satisfy the Big Data characteristics for Big Data classification. To alleviate this problem this paper suggests a new definition for Big Data by introducing a 3D space, C3, which is defined based on three new parameters, cardinality, continuity, and complexity. This paper suggested an integration of modern technologies, Hadoop Distributed File Systems and Cloud Technologies, with the latest representation-learning technique and support vector machine to predict network intrusions through Big Data classification strategy. The cardinality continuity and complexity parameters add extra difficulties to the task of managing the Big Data. Hence the network topology must be designed in such a way that the Big Data Analytics problem can be handled efficiently with cost effectiveness objectives. The challenge here is to minimize the communication cost while satisfying the additional storage and data requirement from public cloud for processing Big Data.

### *G. Anomaly teletraffic intrusion detection systems on hadoop-based platforms*

Jeong et al. [7] give an overview of issues encountered with Intrusion Detection and Big Data and how various Hadoop technologies can address these challenges, specifically focusing on anomaly-based (misuse) IDSs. They describe various techniques and issues found with Intrusion Detection, as well as what some of the main issues are in applying Hadoop technologies for Intrusion Detection.

### *H. H.ELM based Intrusion Detection Algorithm*

Junlong Xiang et al. [10]  introduced a a machine learning method which scales horizontally without losing detection accuracy. A recent advance in a machine learning algorithm called Extreme Learning Machine (ELM) for the MapReduce programming model is proposed. ELM  outperformed Support Vector Machine (SVM) in terms of accuracy, training speed and userfriendliness (automation) improved detection efficiency achieved by increasing the training data used during the learning process.  The solution is computationally heavy. However, due to horizontal scaling through the MapReduce programming model, machine learning based intrusion detection using ELM can extend its applicability to significantly larger datasets than datasets currently used in most papers. This is possible without increasing training time drastically, due to the near linear scaling ability of the proposed ELM algorithm.

## VI. CONCLUSION

From a security perspective, the major concerns of Big Data are privacy, integrity, availability, and confidentiality with respect to outsourced data. As the use of Big data has increased, the security is very important  therefore, the intrusion an important feature for the deployment of Big Data environment detection systems are brought into consideration. This paper summarizes security threat, intrusion detection techniques in Big Data and also an attempt has been made to explore the security mechanism widely used to handle those attacks. Recent research findings uniting IDS specifically in Big Data have been discussed. Almost all of designed algorithms try to detect attacks in Big Data but it appears that more work must be done in the field of Big Data. This survey will hopefully motivate future researchers to come up with smarter and more robust security mechanisms and make their network safer.

## REFERENCES

[1] Chun-Wei Tsai, Chin-Feng Lai, HanChieh Chao  Athanasios V. Vasilakos, Big data analytics: a survey,Journal of BigData,oct2015 .

[2] Shaik Akbar *, T. Srinivasa Rao , Mohammed Ali Hussain , “A Hybrid Scheme based on Big Data Analytics using Intrusion Detection System”,IJST,Vol9,Issue33.

[3] Harshawardhan S. Bhosale,Prof. Devendra P,” Gadekar,” A Review  Paper on Big Data and Hadoop “,International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014

[4] Suthaharan S (2013) Big data classification: problems and challenges in network intrusion prediction with machinelearning. In: Big Data Analytics Workshop, in Conjunction with ACM Sigmetrics. ACM, Pittsburgh, PA, US

[5] Group BDW big data analytics for security intelligence. Available from:

http://downlods.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_intelligence.pdf

[6] Sonal Ashok Hajare", Detection of Network Attacks Using Big Data Analysis", IJRITCC,Vol4,Issue5,2016,pp86-88

[7] Jeong H, Hyun W, Lim J, You I (2012) Anomaly teletraffic intrusion detection systems on hadoop-based platforms:

[8] A survey of some problems and solutions. In: Network-Based Information Systems (NBiS), 2012 15th internationalconference on. IEEE, Melbourne, Australia. pp 766–770

[9] PrathibhaPD ,Dillesh ED,"Design of a Hybrid intrusion detection system using Snort and Hadoop",IJCA,Volume73,No.10,July2013

[10] JunlongXiang,MagnusWesterland,DusanSovilj,GoranPulkkis"Using extreme learning machine for intrusion detection in a big data environment" Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop

[11] Jingwei Huang, Zbigniew Kalbarczyk, and David M. Nicol," "Knowledge Discovery from Big Data for Intrusion Detection Using LDA" 2014 IEEE International Congress on Big Data

[12] Sanjai Veetil,Qigang Gao," A Realtime Intrusion Detection System by Integrating Hadoop and Naive Bayes Classification" DCSI 2013.