IJCTA, 9(22), 2016, pp. 233-239

Anomaly Detection in Network Traffic from Large Dataset

Aparna Bali * Sanjay Madan Rakesh Kumar Sehgal*** Neeraj Sharma******

Abstract : Analyzing network traffic of an organizational network is difficult in real time for the detection of anomalies as the network traffic data volume is very large. Some of the commercial tools for the analysis of network traffic are available like wireshark, tcpdump, packetsquare but they are useful when the amount of data to be analyzed is very less and not in real time. Scalability is the major concern when data to be analyzed is increased to terabytes or petabytes making it difficult to analyze, because a large data set necessitates high computing and storage resources. In this paper, we propose a framework to analyze large amount of network traffic data to segregate anomalies for malicious behavior by using big data frameworks. Here presented the evaluations based upon the 4 months dataset obtained from packet capture files collected through honeypot deployed with multiple vulnerabilities since Jan, 2016 onward. The model has been setup by preprocessing of the dataset using Apache Pig then analyzed the processed data through IDS to build up the machine learn model using Apache Mahout to classify the dataset and validated the findings through the already classified dataset with 93% accuracy.

Keywords : Network Traffic Analysis, IDS, Machine Learning, Big Data.

1. INTRODUCTION

Intrusion is basically a set of activities which are intended to compromise the security of a system and its network components in terms of confidentiality, integrity and availability [1]. Intrusion Detection plays a major role for identifying security breaches after their occurrences. IDSs are capable of spotting and generating alerts for possible cyber-attacks which are categorised as signature based IDS and anomaly based IDS. A signature based IDS consists of a database of known attacks and compares incoming traffic with the existing signatures in the database to verify attack matches. However they have a Limitation that any novel attacks cannot be detected. Whereas anomaly based IDS is based on the network behaviour of finding exceptions in the network traffic that do not match with the normal behaviour. If the network behaviour is not in agreement with the predefined behaviour, then it is considered anomalous. Further there are two classes of network anomalies, performance associated anomalies and security associated anomalies. Broadcast outburst, temporary congestion, gabbling node and server crash are considered as performance associated anomaly on the other hand security related anomalies are caused as a result of malicious actions of the intruder which intentionally flood the network to cause congestion and ultimately leading to disruption of services provided to legitimate users. The capacity of processing, analysing and evaluating network traffic data and to recognize anomalous patterns is increasing rapidly because as the number of internet users are increasing rapidly so is the data being generated by them. There is a need to develop a flexible, fault-tolerant and scalable system to analyse network traffic for anomalies at a large scale. This can be done by developing a framework which is capable of detecting anomalies in large dataset by using machine learning algorithms [2].

* Department of Computer Science & Engineering Chandigarh University, Mohali-140413, India
** Cyber Security Technology Division Center for Development of Advanced Computing, Mohali-160071, India
**** Cyber Security Technology Division Center for Development of Advanced Computing, Mohali-160071, India
**** Department of Computer Science & Engineering Chandigarh University, Mohali-140413, India

The very famous DARPA and KDD datasets have contributed greatly in the intrusion detection field but they lack precision and capability to imitate real world situations. There is a need for novel and dynamic datasets which are capable of reflecting intrusions and latest traffic patterns. On the contrary the datasets used widely by researchers even today are obsolete, inflexible and irreproducible [7]. To conquer these limitations, a new evaluation dataset is used which is obtained from packet capture files collected by honeypot, built on the 4 months of real network traffic data collected from Jan. 2016 to Apr. 2016. There are number of tools like wireshark, tcpdump, packetsquare available to analyze network traffic. These tools are useful when the amount of data to be analyzed is small. Scalability Issue arises when data to be analyzed is increased to terabytes or petabytes. It is difficult to analyze such large amount of data because a large data set necessitates high computing and storage resources. Here big data frameworks like apache pig and apache hive comes to picture which can analyse and query large network traffic easily solving the scalability issues.

The composition of this paper is structured in following sections. The introduction and broad-spectrum overview of large scale network traffic analysis is presented in Section I. Section II covers the literature study following the Section III which presented the methodology of proposed framework for network traffic analysis and anomaly detection. Experimental results are presented in Section IV. Concluding remarks and future work are given in Section V.

2. RELATED WORKS

M.Roesch et al. [18] proposed an intrusion detection system, Snort, which is a signature, based IDS and has a database of stored signatures and rules which match the network traffic for any known signature. If a match occurs alerts are generated and stored in an alert file. The limitation of IDS is that it is not capable of detecting any novel attack.

Sundaram Aurobindo [21] presented an overview of intrusion detection systems. He described the types of IDS, anomaly based and signature based. Further he explained about models and directions in research towards intrusion detection. He concluded that a hybrid model made from both anomaly based and signature based IDS can be used in future. According to the author neither of the models was capable to detect all the intrusions on its own.

Prathibha.P et al. [3] discussed that when large scale network traffic is to be analysed, there are challenges of managing a huge quantity of data for the processing. The work proposed a technique through which the packets acquired by Snort are analyzed using Hadoop. Snort is an intrusion detection tool which is signature based intended to sniff real-time traffic for detecting suspicious anomalies.

Ibrahim LT et al. [19] expressed concern over data explosion issues. Many organizations have to face the challenge to handle, capture, and monitor the data due to the ever-increasing volume of data sets, varying from quite a few terabytes to manifold petabytes. The paper proposed a method to resolve the issue of processing the large dataset by introducing a traffic supervising system based on hadoop which performs analysis of network traffic at a large scale.

Uriel Carrasquilla et al. [4] presented the anomaly detection algorithms of both scattered and clustered anomalies in large dataset based upon space and time complexity. They concluded that there are certain constraints during analysis for the detection of anomalies in the available large datasets like memory limitation, various attributes and big file size. The machine learning tool WEKA, is deficient in its memory management due to its necessity to keep the whole dataset in its memory making it unsuitable for large scale data analysis.

Rong.C et al. [6] compared K-means and Fuzzy c-means algorithms for clustering of a big data set using Mahout. Mahout is one of the frameworks that run on the top of hadoop system which offers scalable machine learning algorithms making it extremely useful for large datasets.

Ali Shiravi et al. [7] presented concern over the availability of suitable dataset in the networking domain. As network traffic patterns and behaviour vary and intrusions evolve, it is a necessity to shift from static and decade old datasets toward the datasets which expose current traffic composition and intrusions. Numerous datasets

which are used for research purpose are internal to the organisations and their sharing is not possible due to privacy concerns whereas other publically available datasets are deeply vague and random which fail to reveal current network trends and traits.

Harneet kaur et al. [16] proposed a tool UAC, URL Analyzer and Classifier for identification of malicious web pages which worked by DOM parsing and javascript analysis.

Michael Baker et al. [15] proposed a tool for network traffic analysis using Packetpig. It is a network security monitoring tool which uses Apache Hadoop for storing enormous full packet captures and Apache Pig for data analysis due to its simplicity of programming. It has integrated Snort IDS and is capable of detecting intrusions based on signatures. The packet captures are stocked up in a computer cluster where MapReduce jobs are initiated by the user using Apache Pig. It is made up of pig's user defined functions, open source tools, pig loaders which allow analyses of large data set. The data created by running pig queries can be visualised using R statistical programming.

Anjali P.P et al. [13] did a survey on broad scope of Apache Pig framework. The key features of the network flows were extracted for data analysis using available tools for packet capture. The flow analyzer based on Pig was implemented and traffic flow analysis was done easily without much programming skills. It reduced the time consumption greatly by reducing the complex programming constructs required for a MapReduce concept using Java programs. Programming with Pig has the advantages of Map/Reduce jobs because of the layered design and built in compilers.

Shan Suthaharan et al. [22] focused on challenges of classification of big data for intrusion detection and utilization of machine learning techniques of big data for intrusion detection. Different types of supervised and unsupervised learning techniques are discussed and various big data technologies like hadoop, hive for classification problem of network traffic are also used.

3. PROPOSED FRAMEWORK FOR DETECTION OF ANOMALIES IN LARGE DATASET

3.1. Network Anomaly Detection

Anomaly is something which is not normal and do not match with expected normal behavior in the network. For instance, in a network, anomalous traffic pattern could mean a compromised system sending sensitive information to an unauthorized system. There exist three broad categories of anomaly detection techniques, Unsupervised, Supervised and Semi-supervised. Unsupervised detection techniques identify anomalies in unlabelled data set by making assumption that in a dataset frequency of normal instances are more than abnormal instances and if assumption is wrong it leads to high false alarm rates. Supervised detection techniques identify anomalies in labeled data set and a model is built to classify normal and abnormal class. To determine from which class a new data instance belong, it is compared against the already built model. In Semi-supervised detection techniques only normal instances in the dataset are labeled and rest data set is assumed abnormal. In the proposed work we followed supervised anomaly detection technique. Using apache pig, dataset is analyzed and relevant features are extracted. For anomaly detection, the dataset is examined using a signature based IDS and the alerts generated by the IDS are used for labeling the dataset instances as malicious or benign. For building a ML model the dataset is divided into training and testing set. Apache mahout is used to build a model using random forest algorithm and the large dataset can be classified based upon machine learn model. The experimental results provide information about recent cyber attacks captured in our honeypots. The IDS is successful in identifying potentially bad traffic, several Misc attacks, web application attacks, network Trojans, denial of service attacks, network scans and several executable codes. These observations are useful to understand characteristics and trends of latest cyber threats and to develop counter steps against them. Figure 1 shows the proposed framework for the detection of anomaly from large dataset.

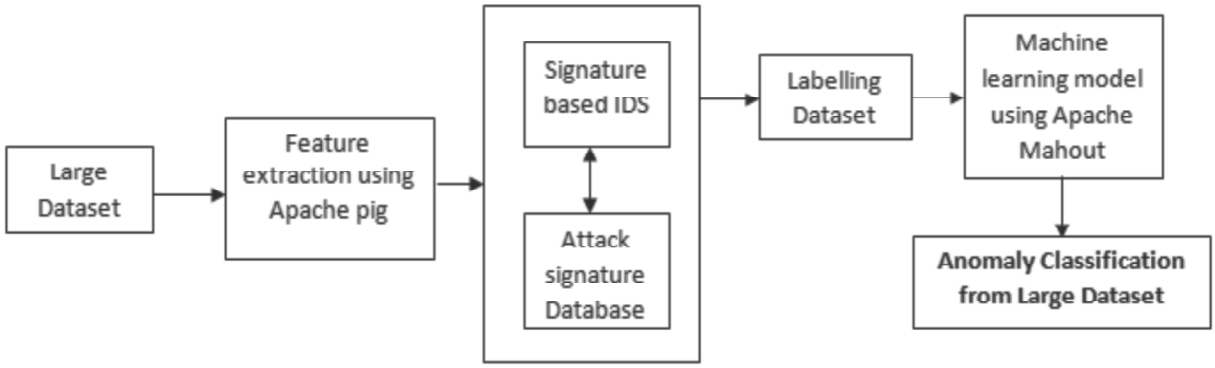

Fig. 1. Anomaly Detection approach.

3.2. Dataset

For experiments and implementation, we used a dataset obtained from raw packet captures collected by honeypot, built on the 4 months of real network traffic data collected from Jan 2016 to Apr 2016. Therefore the dataset has approximately 10 lakhs entries out of which 2 lakhs are malicious. Every entry of data set consists of 29 attributes namely IP version, IP header, time of service, total length, identification, IP flags, flag offset, time to live, protocol, checksum, source address, destination address, TCP source port, TCP destination port, TCP offset, TCP nonce sum, TCP congestion window reduced, TCP ECN-echo, TCP urgent, TCP acknowledgement, TCP push, TCP reset, TCP syn, TCP fin, TCP windows, TCP length, UDP source port, UDP destination port and UDP length. Further for validation purpose, three more datasets collected on 3 consecutive days 1 June 2016 to 3 June 2016 from the honeypot were used.

3.3. Analysing Data using Apache Pig

Apache pig [11] is a framework for analysis of enormous sets of data demonstrating them as data flows. In Hadoop using Apache Pig, data manipulation operations can be performed. A high-level language pig Latin is used for scripting the data analysis programs. All the Pig Latin scripts are internally transformed to Map and Reduce tasks. The Pig Engine processes the Pig Latin scripts as input and transforms these scripts to MapReduce jobs. In the proposed work, feature extraction is performed using Packetpig [14], a network security monitoring tool which uses Apache Hadoop for storing enormous full packet captures and Apache Pig for data analysis. The packet captures are stocked up in a computer cluster where MapReduce jobs are initiated by the user using Apache Pig.

3.4. Machine learning using Apache Mahout

Weka is an esteemed, well-built project for building machine learning model from datasets. It comprise of a vast set of well optimized machine learning algorithms but its limitation is that for memory intensive tasks on large datasets, it's package is not well optimized. Apache Mahout [6] is one of the frameworks that run on the top of hadoop system which offers scalable machine learning algorithms such as clustering, classification, logical regression, recommendations, dimension reduction etc. In machine learning systems, the accuracy of the system built depends on the size of data used, indicating that as the amount of data used for training is increased and so is the Mahout's performance. In the proposed work, Random Forest algorithm is used built on mahout for training the dataset and then used it with test dataset to generate the confusion matrix for the performance evaluation. The confusion matrix shows correctly classified instances, incorrectly classified instances, total classified instances, accuracy and reliability of the classified dataset**.**

3.5. Classification of Dataset

For the classification of data instance as malicious or benign, a label is associated with it. It is a difficult job to correctly label data as normal or anomalous. Labelling is done by data experts manually hence requiring significant effort to achieve the labelled training dataset. Suricata [16] is a rule-based Intrusion detection system that uses

superficially made rule sets to examine network traffic and generate alerts when any suspicious event takes place. The packet capture dataset was analysed using suricata IDS, It generated alerts about the malicious data. The labelling of dataset was done according to suricata analysis results.

4. EXPERIMENT AND RESULT

In the proposed work, we used random forests for classification as they are most accurate amongst present algorithms and efficient on large datasets as they are able to successfully classify enormous data with accuracy. Adele Cutler and Leo Breiman developed Random forest algorithm which is an ensemble learn technique for regression, classification and many other tasks. They work by constructing multiple decision trees during training period and as an output the mode of the classification or mean prediction about regression of individual trees is depicted. It is a flexible machine learning algorithm capable of accomplishing both classification and regression tasks. It can hold large number of input variables from which the most significant variables are identified so it serves the purpose of dimensionality reduction methods as well. They have an effective way to estimate missing data when a huge fraction of data is missing. They also balance errors in datasets in case of imbalanced classes. They consist of a collection of regression trees which are relevant for intrusion detection and spawn many categorization trees. In random forest, multiple trees are grown in contrast to a single tree as in CART model. New objects are classified based on attributes, each tree choose a class by voting for it. The classification which has majority votes is selected over every tree in the forest.

4.1. Building a prediction Model using Random-forest Algorithm

The Results of Machine learning algorithm Random forest which is built on apache mahout are discussed in this section. For large scale analysis we transferred the classified dataset to hadoop distributed file system. Before building a model using random forest, a descriptor file is generated which has information about classified dataset. Then random forest model is built using 100 trees on the basis of training dataset. Once the data is trained, the model is used to classify new dataset. The results of training dataset gave a confusion matrix which depicted 146117 instances were correctly classified and 10681 instances were incorrectly classified as anomalies. The accuracy percentage is 93.18 (correctly classified instances by the model / classified instances). Thus a system is built for anomaly detection.

Table 1 shows the correctly classified instances, incorrectly classified instances, total classified instances and accuracy of the dataset used to build the model. To validate the model, we took 3 more datasets collected on 3 consecutive days 1 June 2016 to 3 June 2016 from the honeypot server.

Table 1. Experiment Result

Title	*Jan-April 2016*	*1 June 2016*	*2 June 2016*	*3 June 2016*
Correctly Classified Instances	146117	87184	151902	158986
Incorrectly Classified Instances	10681	15138	15433	16536
Total Classified Instances	156798	102312	167337	175522
Accuracy	93.18%	85.21%	90.07%	90.57%

5. CONCLUSION

In this paper, proposed a framework for the identification of anomalous behavior in large network traffic obtained from raw packet captures by honeypot Server, built on the 4 months of real network traffic data collected from Jan. 2016 to Apr. 2016. The honeypots have succeeded in capturing various types of attacks and botnets

which are very effective to understand the trends and traits of recent cyber threats thereby also helping to device counter measure to mitigate them. We utilize Hadoop based frameworks for large-scale log analysis. Our dataset was divided into training data and testing data. A machine learning model was built on top of hadoop using apache mahout. The model was trained with the training dataset then it was investigated for accuracy using testing dataset. The main objective of the work is to proficiently classify the large set of data as malicious or benign based on the Apache Mahout/Hadoop framework through finding anomalous behavior in dataset. The classification was done using random forest algorithm. The results showed that the model trained using random forest algorithm gives 93.18 percent correct classifications.

6. ACKNOWLEDGMENT

This work has been carried out at Cyber Security Technology Division, C-DAC, Mohali and we are thankful to the organization and director, Sh. D.K. Jain for his support and team members for their time to time guidance.

7. REFERENCES

1. Bhuyan MH, Bhattacharyya DK, Kalita JK.. Network anomaly detection: methods, systems and tools. IEEE Communications Surveys & Tutorials. 2014 Feb 7;16(1):303-36.
2. Zuech R, Khoshgoftaar TM, Wald R. Intrusion detection and big heterogeneous data: a survey. Journal of Big Data. 2015 Feb 27;2(1):1.
3. Prathibha PG, Dileesh ED. Design of a hybrid intrusion detection system using snort and hadoop. International Journal of Computer Applications. 2013 Jan 1;73(10):5-10.
4. Carrasquilla U. Benchmarking algorithms for detecting anomalies in large datasets. MeasureIT, 2010 Nov;1-6.
5. Athira AB, Pathari V. Standardisation And Classification Of Alerts Generated By Intrusion Detection Systems. International Journal on Cybernetics & Informatics (IJCI).2016 April 2; 5(2):21-29.
6. Rong C. Using Mahout for clustering Wikipedia's latest articles: a comparison between K-means and fuzzy C-means in the cloud. InCloud Computing Technology and Science (CloudCom), IEEE Third International Conference.2011 Nov 29;565-569.
7. Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security. 2012 May 31;31(3):357-74.
8. Song J, Takakura H, Okabe Y, Eto M, Inoue D, Nakao K. Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. InProceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security.2011 Apr 10;29-36.
9. Therdphapiyanak J, Piromsopa K. An analysis of suitable parameters for efficiently applying K-means clustering to large TCPdump data set using Hadoop framework. InElectrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference.2013 May 15;1-6.
10. Ertoz L, Eilertson E, Lazarevic A, Tan PN, Kumar V, Srivastava J, Dokas P. MINDS-minnesota intrusion detection system. Next generation data mining. 2004 Oct 1:199-218.
11. Olston C, Reed B, Srivastava U, Kumar R, Tomkins A. Pig latin: a not-so-foreign language for data processing. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008 Jun 9;1099-1110.
12. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Zhang N, Antony S, Liu H, Murthy R. Hive-a petabyte scale data warehouse using hadoop. IEEE 26th International Conference on Data Engineering (ICDE).2010 Mar 1; 996-1005.
13. Anjali PP, Binu A. Network Traffic Analysis: Hadoop Pig vs Typical MapReduce. arXiv preprint arXiv. 2013 Dec 19; 1312-5469.
14. Gupta A. Apache Mahout Clustering Designs. Packt Publishing Ltd;Oct 2015.
15. About packetpig, Baker M, Turnbull D, Kaszuba G. Finding needles in haystacks(the size of countries) http://docs.huihoo.com/blackhat/europe-2012/bh-eu-12 BakerNeedles_Haystacks-WP.pdf.
16. H. Kaur, S. Madan, RK. Sehgal, "UAC: A Lightweight and Scalable Approach to Detect Malicious Web Pages", Modern Trends and Techniques in Computer Science: 3rd Computer Science On-line Conference 2014 (CSOC 2014), Springer International Publishing, pp. 241-261.

17. About Suricata , https://suricata-ids.org/ Date accessed: 05/04/2016.

18. Roesch M. Snort: Lightweight Intrusion Detection for Networks. In LISA.1999 Nov 7; 229-238.

19. Ibrahim LT, Hassan R, Ahmad K, Asat AN. A study on improvement of internet traffic measurement and analysis using Hadoop system. InElectrical Engineering and Informatics (ICEEI), 2015 International Conference.2015 Aug 10; 462-466.

20. Therdphapiyanak, Jakrarin, and Krerk Piromsopa. "An analysis of suitable parameters for efficiently applying K-means clustering to large TCPdump data set using Hadoop framework." In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference on, pp. 1-6. IEEE, 2013.

21. A. Sundaram, "An introduction to intrusion detection," Crossroads, vol. 2, no. 4, pp. 3–7, April 1996.

22. Suthaharan, Shan. "Big data classification: Problems and challenges in network intrusion prediction with machine learning." *ACM SIGMETRICS Performance Evaluation Review* 41, no. 4 (2014): 70-73.