Big Data Analytics for Network Anomaly Detection from Netflow Data

Duygu Sinanc Terzi
Gazi University
Computer Engineering
Ankara, Turkey
duygusinanc@gazi.edu.tr

Ramazan Terzi
Gazi University
Computer Engineering
Ankara, Turkey
ramazanterzi@gazi.edu.tr

Seref Sagiroglu
Gazi University
Computer Engineering
Ankara, Turkey
ss@gazi.edu.tr

***Abstract*— Cyber-attacks wasorganized in a simple and random way in the past. However attacks are carried out systematically and long term nowadays. In addition, the high calculation volume and continuous changes in network data distribution have made it more difficult to analyze data and detect abnormal behaviors within. For this reason, big data solutions have become essential. In this paper, firstly network anomaly and attack detection studies on big data has been reviewed. Then, a public big network data was analyzed with a new unsupervised anomaly detection approach on Apache Spark cluster in Azure HDInsight. Finally, the results obtained from a case study were evaluated, %96 accuracy was achieved. The results were visualized after dimension reduction using Principal Component Analysis (PCA). The identified anomalies may provide usable outputs to understand the behavior of the network, distinguishing the attacks, providing better cyber security, and protecting critical infrastructures.**

Keywords—network anomaly detection, network behaviour analysis, big data security analysis, big data, netflow, UDP DDoS

I. INTRODUCTION

Cyber security is becoming increasingly important; therefore, countries have started to make big investments in order to protect their critical infrastructures. President's Fiscal Year 2016, budget request for US Department of State Security Network Security Distribution Department was $479.8 million [1]. According to Norton, cybercrime victims have spent $126 billion globally since 2015 [2]. The reason for these huge cybersecurity investments is that cybercrimes are becoming more and more intelligent, complex, and destructive.

Conventional defense systems are inefficient because they mostly cannot detect these attacks because of their signature-based structure and it is difficult to carry out both operations and analysis of huge amount of security data simultaneously. Thus, Security Information and Event Management (SIEM) systems are now being replaced by Big Data Security Analytics systems [3]. Big Data Security Analytics is gaining great importance as records do not have to be deleted after a certain period of time, complex queries can be answered in a short time, non-structural data can be analyzed easily, and cluster computing infrastructures are increased reliability [3]. For this reason, it has great importance to examine the behavioral and statistical changes on big network data to determine anomalies and attacks with or without signatures.

One of the intelligent solutions that companies use to protect their networks from emerging threats is to collect IP traffic flows and deploy anomaly detection systems based on network traffic monitoring [4]. In the direction of network anomaly and intrusion detection studies on big data, a new unsupervised anomaly detection approach has been proposed in this paper. It is aimed to determine the anomalies caused by the UDP flood attack from specific IPs. This approach is implemented on a public NetFlow data within a case study.

II. RELATED WORKS

Big data is a massive data collection that includes different and diverse type of datasets [5]. Big data characteristics defined by V's, generally 6V's [6, 7] as seen in Table I. Velocity refers to the speed of processing and creation of data. Volume is the amount of data. Variety indicates the types of data. Veracity points to the trustworthiness of data. Vocabulary involves schema, models, and ontologies that describe the data's structure. Value refers to insight and cost. Because traditional methods cannot cope with the characteristics of big data, big data analytics is gaining importance. Big data analytics is a set of well-established tools and techniques to find useful hidden information inside the raw data [8]. Hence, big data models are more rapid, scalable, and fault-tolerant than traditional approaches.

TABLE I. COMPONENTS OF BIG DATA

Velocity	Speed of data
Volume	Size of data
Variety	Diversity of data
Veracity	Uncertainty of data
Vocabulary	Data about the data
Value	Usefulness of data

A threat or an intrusion attempt refers to create an anomaly as unauthorized try to access system, alter information, or make system unusable [9]. Anomaly detection approaches are used in many applications such as intrusion detection, fraud detection, and data leakage prevention [10]. Network-based intrusion detection is purposed to detect unusual behavior patterns in of network users, and the high speed of the interfaces required big

978-1-5386-0930-9/17/$31.00 ©2017 IEEE

data analytics for this process as a natural result of the development.

When examining the literature in terms of big network anomaly detection (Table II), it seems that the network analysis with big data is mostly carried out by conventional methods on relatively high volume data to identify attacks via supervised techniques. In addition, big data solutions have begun to be introduced to reduce false positive and false negative rate, and to handle huge and stream data.

III. Proposed Approach

NetFlow is a network protocol that collects traffic information such as network users, network applications, and routing traffic [26]. This data is widely used for network anomaly detection studies because malicious traffic information can be identified through NetFlow analysis.

Network anomaly detection can be performed by distance based, density based, and machine learning or soft-computing-based approaches [27]. The proposed method in this study is the clustering-based from the machine learning perspective. Clustering is a commonly used method for detecting anomalies as it does not require labeled data sets and pre-defined classes [27].

The steps of the proposed approach are explained below:

1. NetFlows are divided into intervals.
   - Most actions show similar behavior in several minutes (temporal locality behavior [28]).

2. Netflows are aggregated according to source IPs.
   - The data size is reduced for processing.
   - The aggregated data may show new patterns to detect behaviors.

3. The obtained data is standardized by zero score as in (1) where μ is the mean and σ is the standard deviation.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

   - This procedure equalize the data variability.
   - Standardized data is less affected by outlier.

4. The aggregated NetFlows are clustered based on the k-means algorithm as distributed.
   - The unsupervised techniques trained with unlabeled data has the ability to detect unfamiliar attacks [29].
   - It is predicted that clusters will occur according to normal or abnormal traffic behavior.

5. The Euclidean distance of the cluster elements to the cluster center is calculated.
   - The elements in the cluster should be close to the center for a good clustering.
   - The elements may be abnormally distant from the center because of any reason and the centroids can be used for outlier detection.
   - The histogram is used to understand the distribution of distance of the elements from the center.
   - The elements stay distant from the concentrated region on the histogram are considered as anomalous.

6. The actual normal and abnormal flow numbers are determined from time intervals in steps 4 and 5. Finally, the success criterion is evaluated.

IV. Case Study

The proposed approach was implemented on a public NetFlow data on Apache Spark cluster in Azure HDInsight with python programming language. By applying the approach on the case study, 0.96 accuracy rate was obtained. Moreover, the obtained results were visualized as 3D by dimension reduction using PCA.

The CTU-13 dataset was used for botnet traffic analysis. This data was captured by the CTU University, Czech Republic in 2011 [30]. The dataset consists of 13 scenarios having different attack samples. In this study, the 10th scenario was selected because of the size of the dataset and the number of botnet attacks. The data set has 4.75 hours records and 1309791 flows covering 106352 UDP DDoS flows.

```
StartTime,Dur,Proto,SrcAddr,Sport,Dir,DstAddr,Dport,State,sTos,dTos,TotPkts,TotBytes,SrcBytes,Label
2011/08/18 09:56:29.146156,2752.656250,udp,71.222.124.71,60621,  <->,147.32.84.59,63550,CON,0,0,3,435,290,flow=Background-Established-cmpgw-CVUT
2011/08/18 09:56:42.630892,1849.315552,udp,78.234.54.245,51413,  <->,147.32.84.59,63550,CON,0,0,3,417,272,flow=Background-Established-cmpgw-CVUT
2011/08/18 09:56:44.640650,2091.747314,udp,31.147.120.139,63195,  <->,147.32.84.59,63550,CON,0,0,2,290,145,flow=Background-Established-cmpgw-CVUT
2011/08/18 10:10:52.782230,1535.769409,udp,118.5.35.64,39110,  <->,147.32.84.59,63550,CON,0,0,2,290,145,flow=Background-Established-cmpgw-CVUT
2011/08/18 10:19:13.328372,0.002636,tcp,147.32.86.166,33426,  <?>,212.24.150.110,25443,FRPA_FPA,0,0,6,490,321,flow=Background
2011/08/18 10:19:13.328670,72.436790,udp,82.39.2.249,41915,  <->,147.32.84.59,43087,CON,0,0,23849,24298138,509912,flow=Background-Established-cmpgw-CVUT
2011/08/18 10:19:13.330765,3599.473633,tcp,147.32.86.166,42020,  <?>,147.32.192.34,993,PA_PA,0,0,543,98018,33640,flow=Background
2011/08/18 10:19:13.333772,28.152548,tcp,115.184.37.24,49190,  <?>,147.32.84.2,80,A_FPA,0,0,222,191281,6610,flow=Background
2011/08/18 10:19:13.335316,632.001648,tcp,80.78.79.156,51287,  <?>,147.32.86.24,31002,FPA_FPA,0,0,15347,2542390,2237911,flow=Background
2011/08/18 10:19:13.335512,628.915222,udp,147.32.86.24,31002,  <->,151.41.188.39,49621,CON,0,0,9464,2024974,2022958,flow=Background-UDP-Established
2011/08/18 10:19:13.336134,0.000177,udp,82.73.244.56,39051,   ->,147.32.84.118,1153,INT,0,,1,145,145,flow=Background-UDP-Attempt
2011/08/18 10:19:13.336311,0.000000,icmp,147.32.84.118,0x0303,   ->,82.73.244.56,0x8104,URP,0,,1,173,173,flow=Background
2011/08/18 10:19:13.337361,1315.114136,tcp,188.95.61.42,53389,  <?>,147.32.86.110,48190,RPA_FPA,0,0,59405,24973123,21487780,flow=Background
2011/08/18 10:19:13.341725,450.406830,tcp,192.221.106.126,80,  <?>,147.32.84.59,2774,FPA_FA,0,0,85347,97405130,96357830,flow=Background-Established-cmpgw-CVUT
2011/08/18 10:19:13.344200,123.438812,tcp,212.111.2.151,8000,  <?>,147.32.86.135,3978,PA_FRA,0,0,4238,4189542,4104702,flow=Background
```

Fig. 1. The10. scenario of CTU-13 dataset

TABLE II. COMPARISON FOR LITERATURE

Reference	Purpose	Anomalies	Data	Techniques and Technologies	Success Rates / Results
[11]	A big data based model, which can avoid the influence brought by adjustment of network traffic distribution, reduce the false negative rate and increase detection accuracy	Dos, U2R, R2L, Probe	KDD CUP99	k-means, KNN, decision tree, random forest	Detection rates: 95.4% on normal data, 98.6% on DoS attack, 93.9% on Probe attack, 56.1% on U2R attack, and 77.2% on R2L attack
[12]	DDoS detection method implemention in Apache Spark Cluster	DDoS	2000 DARPA LLDOS 1.0 and generated normal traffic data	ANN, Spark	Accuracy: 94%
[13]	A method for analyzing network traffic using Big Data techniques	SYN Flood, NULL scan, XMAS Scan, SYN/FIN Attack	NCCDC	HDFS, Hive	The results are presented visually
[14]	A real-time IDS for ultra-high-speed big data environment using Hadoop	DoS, U2R, R2L, Probing	DARPA, KDD 99, NSL-KDD	J48, REPTree, random forest tree, conjunctive rule, SVM, naïve bayes, Hadoop	REPTree and J48 are the best classifiers in terms of TP: %99.9
[15]	A real-time DDoS attack detection mechanism based on Multivariate Dimensionality Reduction Analysis	DDoS	KDD Cup 1999	Principal Component Analysis, Multivarite Correlation Analysis, MATLAB	The results were presented visually
[16]	An anomaly detection model which combine cloud computing with machine learning based on Hadoop	Bad connections	KDD CUP 99	HDFS, MapReduce, Weka, naïve bayes, decision tree, SVM	Above 90% of accuracy.
[17]	Adaptive stream projected outlier detector to detect anomalies from large datasets using an adaptive subspace analysis	DoS, R2L, U2R, Probing	KDD CUP99 and generated data	Adaptive stream projected outlier detector	The results were presented visually
[18]	A real time hybrid intrusion detection system using apache storm	DDoS	ISCX 2012	Storm, CC4 neural networks, Multi-Layer Perceptron neural networks	The average accuracy: 89%
[19]	Anomaly based intrusion detection at different layers of TCP/IP (network/application)	Abusive internet access, systematic downloading, and DDoS attacks	Proxy server logs of a campus LAN and edge router traces	Machine learning, time series analysis, pattern analysis	The results were presented visually
[20]	A novelty entropy mode for traffic anomaly detection	Dos, DDoS, DRDoS port scan	IPFIX data collected from a universiy's edge router	MapReduce Adjustable Piecewise Shannon Entropy (APSE), Shannon entropy	APSE has better performance than Shannon entropy in traffic anomaly detection
[21]	A live operational and situational awareness implementation based on big data architectures, graph analytics, streaming analytics, and interactive visualizations to a security use case	Advanced Persistent Threats and contextual anomalies	SIEM data from a large Global 500 company	Tableau, MapReduce, Kafka, Apama, GemFireXD,D3.js	Average precision and recall for anomaly groups: very high [0.7,0.9], high [0.75,0.7], medium [0.9,0.9], low [0.8,0.95] and very low [1.0,0.4]
[22]	A MapReduce framework (Hashdoop), that splits traffic with a hash function to detect network anomalies	Sasser, RPC, SMB, Ping, NetBIOS, other attacks	MAWI traffic archive	MapReduce, Cyclic redundancy check hash algorithm	F-score: 0.88
[23]	P2P botnet detection using Random Forests in quasi-real-time	Bot attacks (conficker, kelihos-hlux, zeus,storm, waledac)	CAIDA and campus network traffic	Hive, Tshark, Mahout, Random Forest	Accuracy: 99.7%
[24]	Analyzing Netfow data using Hadoop and evaluating the efficiency of different data formats	Watering hole attack	CAIDA	The data was converted to Hadoop sequence file format. MapReduce jobs were run on the data to detect watering whole attack on Amazon servers using Hive.	The sequence file format is more efficient in Hadoop MapReduce, and the definition of reducer numbers is very important in Hive
[25]	A framework for anomaly detection and forensics in Big Data tackling with the Big Data 4 Vs	Attacks to the DNS/DC, Firewall access attempts FTP attempts to outer nodes, Background IRC activity, Parsing errors	VAST 2012 mini challenge 2	Exponentially Weighted Moving Average , PCA, MEDA, Time lines, oMEDA	The results were presented visually

The implementation steps and the results were explained below:

1. NetFlows in raw data (Fig. 1) were divided into 1 minutes intervals.
   - 1 minute was sufficient to capture anomalies and intervals do not contain too much flows [30].

2. Netflows were aggregated according to the source IPs.
   - The aggregation was carried out according to number of unique source ports, number of unique destination IP addresses, number of unique destination ports, number of NetFlows, number of bytes, and number of packets [30].
   - The flow number was reduced to 1309791 from 294374.

3. The obtained data was standardized by zero score according to their mean and standard deviation values (Table III).

TABLE III. MEAN AND STANDARD DEVIATION VALUES OF DATASET

	Mean	Standard Deviation
Unique Source Ports	3.888	45.826
Unique Destination IP Addresses	1.727	11.0526
Unique Destination Ports	1.529	9.356
Number Of Netflows	4.449	48.753
Number Of Bytes	164432.266	8902618.818
Number Of Packets	215.851	10130.520

4. The aggregated NetFlows were clustered.
   - The algorithm was executed with 2 clusters and max 1000 iterations.
   - After the clustering process, the first cluster had 294351 members and the second cluster had 23 members.
   - Many flows from many ports to few destinations can be considered as botnet behavior. In second cluster, the average unique destination IPs and the average unique destination ports were found to be very small, even though the average unique source ports and the number of NetFlows were very large (Table IV). For this reason, the second cluster was labeled as anomaly.

TABLE IV. FEATURES OF CLUSTERS

	First Cluster	Second Cluster
Total Instances	294351	23
Average Unique Source Ports	-0.0057	72.9266
Average Unique Destination IP Addresses	0	-0.0344
Average Unique Destination Ports	0	-0.1263
Average Number Of Netflows	-0.0054	68.5358
Average Number Of Bytes	0	0.3821
Average Number Of Packets	0	0.309

5. Although the distinction was made by clustering, anomalies can also be found in the normally classified cluster.
   - The distance of the first cluster elements to the first cluster center and the 5-bucked histogram of cluster was calculated.
   - The distant elements from the cluster center were considered as anomalies.

```
[0.05707889422774315, 62.087911650538445,
124.11874440684915, 186.14957716315985,
248.18040991947055, 310.21124267578125],
[294318, 22, 6, 3, 2]
```

Fig. 2. 5-bucked histogram of first cluster

   - According to the tuple ([distance], [number of elements]) as seen Fig. 2, all elements far from the first intense distance were considered as anomalies. There are 33 (22, 6, 3, 2) elements away from 0.0570… .

6. The aggregated flows were determined by the time intervals to actually cover how many flows in the raw data.
   - In order to find the success rate, IPs that attacked the botnet within the scenario and "botnet" in the flow label were evaluated as anomalies.
   - It was determined that the cluster identified as the anomaly in step 4 was actually anomalous. However, none of the cluster elements identified as anomaly in step 5 was anomalous. This deviation may be an outlier due to any reason.
   - The number of times the aggregated flows in the raw data was calculated. These are respectively 76954 and 15458 in step 4 and 5.
   - The confusion matrix was obtained as a result of the analysis (Table V).

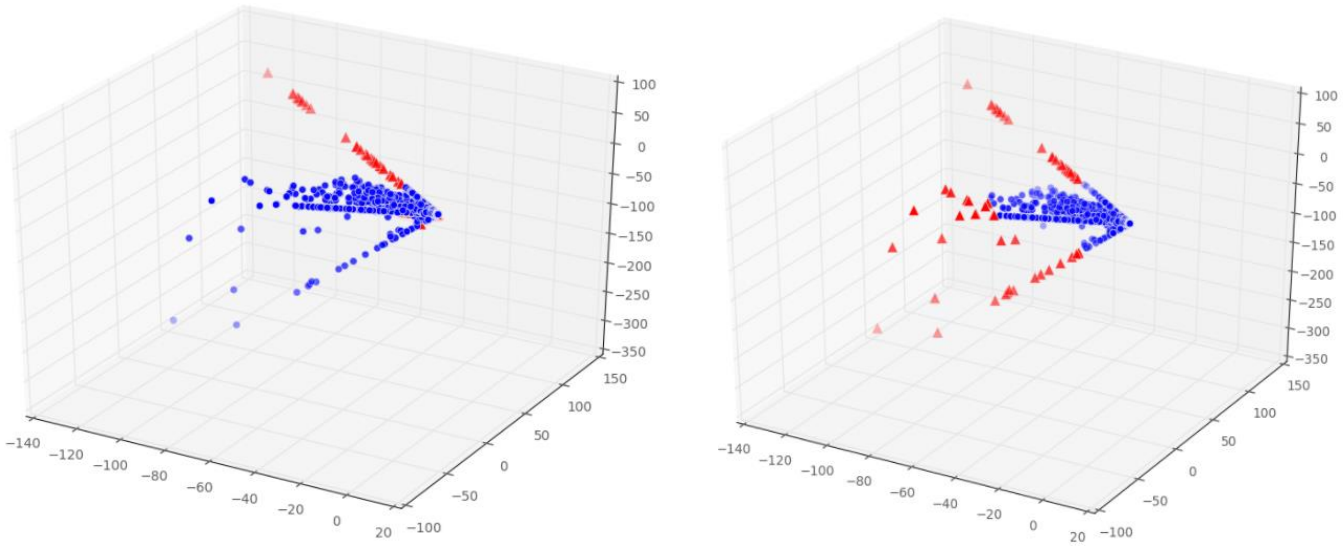
TABLE V. CONFUSION MATRIX

	Actually Botnet	Actually Not Botnet
Detected Botnet	TP=76954	FP=15458
Detected Not Botnet	FN=29398	TN=1187981

   - According to the confusion matrix, the accuracy (2) of the unsupervised anomaly detection approach is 0.96.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

   - Besides the accuracy, PCA was used to make the data easy to explore and visualize. The 6-dimensional data was reduced to 3-dimensions with PCA. Red triangles represent botnet traffic, and blue circles represent normal traffic in Fig. 3.



(a) (b)

Fig. 3. (a) Anomalies in data, (b) Anomalies were detected by the proposed method

- The proposed method was able to detect anomalies that were different from normal successfully. However cannot detect anomalies that were similar to normal traffic. Somehow, a group of data that had a different pattern from normal were detected as abnormal incorrectly.

V. Conclusion

In this paper, a public data was analyzed with a new unsupervised anomaly detection approach on Apache Spark cluster in Azure HDInsight with 96% accuracy. The obtained results were visualized as 3D by dimension reduction with PCA. By this way, suspicious or malicious traffic flows, outliers, compromised devices, and policy violations were detected easily.

The results and the literature clearly point out that timely and effectively detecting anomalies are necessary for better network security. The identified anomalies may provide better perceptions to distinguish, analyze, and understand. High accuracy in anomaly detection provides high quality of services and communication even if the complexity of attacks and analysis process are increased.

In network traffic, most of the flows are normal. Anomalies like attacks and outliers are naturally rare. It is a situation that negatively affects the detection of anomalies and the success rates. For this reason, more and better results might be achieved in future studies having more data and anomalies, innovative algorithms and platforms. These issues are considered to be focused on in future studies.

References

1. *Budget-in-Brief Fiscal Year 2016*, US Department of Homeland Security, Editor. 2016.
2. *2016 Norton Cyber Security Insights Report*. 2016.
3. Big Data Working Group, *Big Data Analytics for Security Intelligence*. 2013, Cloud Security Alliance.
4. Cisco Public, *Network as a Security Sensor Threat Defense with Full NetFlow*. 2016. 1-19.
5. Elarabi, T., et al. *Big data analytics concepts and management techniques*. in *2016 International Conference on Inventive Computation Technologies (ICICT)*. 2016.
6. Lakshen, G.A., S. Vraneš, and V. Janev. *Big data and quality: A literature review*. in *2016 24th Telecommunications Forum (TELFOR)*. 2016.
7. Tsai, C.-W., et al., *Big data analytics: a survey.* Journal of Big Data, 2015. 2(1): p. 21.
8. Sanjay, M. and B.H. Alamma. *An insight into big data analytics; Methods and application*. in *2016 International Conference on Inventive Computation Technologies (ICICT)*. 2016.
9. Bhuyan, M.H., D.K. Bhattacharyya, and J.K. Kalita, *Network Anomaly Detection: Methods, Systems and Tools.* IEEE Communications Surveys & Tutorials, 2014. 16(1): p. 303-336.
10. Goldstein, M. and S. Uchida, *A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data.* PLOS ONE, 2016. 11(4): p. e0152173.
11. Yao, H., Y. Liu, and C. Fang, *An Abnormal Network Traffic Detection Algorithm Based on Big Data Analysis.* International Journal of Computers, Communications & Control, 2016. 11(4).
12. Hsieh, C.-J. and T.-Y. Chan. *Detection DDoS attacks based on neural-network using Apache Spark*. in *Applied System Innovation (ICASI), 2016 International Conference on*. 2016.
13. Bachupally, Y.R., X. Yuan, and K. Roy. *Network security analysis using Big Data technology*. in *SoutheastCon, 2016*. 2016.
14. Rathore, M.M., A. Ahmad, and A. Paul, *Real time intrusion detection system for ultra-high-speed big data environments.* The Journal of Supercomputing, 2016. 72(9): p. 3489-3510.
15. Jia, B., et al., *A Novel Real-Time DDoS Attack Detection Mechanism Based on MDRA Algorithm in Big Data.* Mathematical Problems in Engineering, 2016. 2016.
16. Cui, B. and S. He. *Anomaly Detection Model Based on Hadoop Platform and Weka Interface*. in *2016 10th International Conference on*

Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS). 2016.
17. Zhang, J., et al., *Detecting anomalies from big network traffic data using an adaptive detection approach.* Information Sciences, 2015. 318: p. 91-110.
18. Mylavarapu, G., J. Thomas, and A.K. TK. *Real-time Hybrid Intrusion Detection System using Apache Storm*. in *High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on*. 2015.
19. Sait, S.Y., et al., *Multi-level anomaly detection: Relevance of big data analytics in networks.* Sadhana, 2015. 40(6): p. 1737-1767.
20. Tian, G., et al. *Mining network traffic anomaly based on adjustable piecewise entropy*. in *Quality of Service (IWQoS), 2015 IEEE 23rd International Symposium on*. 2015.
21. Puri, C. and C. Dukatz. *Analyzing and Predicting Security Event Anomalies: Lessons Learned from a Large Enterprise Big Data Streaming Analytics Deployment*. in *Database and Expert Systems Applications (DEXA), 2015 26th International Workshop on*. 2015.
22. Fontugne, R., J. Mazel, and K. Fukuda. *Hashdoop: A MapReduce framework for network anomaly detection*. in *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*. 2014.
23. Singh, K., et al., *Big data analytics framework for peer-to-peer botnet detection using random forests.* Information Sciences, 2014. 278: p. 488-497.
24. Zhou, X., et al. *Exploring Netfow data using hadoop*. in *Proceedings of the Second ASE International Conference on Big Data Science and Computing*. 2014.
25. Camacho, J., et al. *Tackling the Big Data 4 vs for anomaly detection*. in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. 2014.
26. He, W., G. Hu, and Y. Zhou, *Large-scale IP network behavior anomaly detection and identification using substructure-based approach and multivariate time series mining.* Telecommunication Systems, 2012. 50(1): p. 1-13.
27. Liu, D., et al. *Network Traffic Anomaly Detection Using Adaptive Density-Based Fuzzy Clustering*. in *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*. 2014.
28. Xie, Y., et al., *Resisting Web Proxy-Based HTTP Attacks by Temporal and Spatial Locality Behavior.* IEEE Transactions on Parallel and Distributed Systems, 2013. 24(7): p. 1401-1410.
29. Gogoi, P., B. Borah, and D.K. Bhattacharyya, *Anomaly detection analysis of intrusion data using supervised & unsupervised approach.* Journal of Convergence Information Technology, 2010. 5(1): p. 95-110.
30. García, S., et al., *An empirical comparison of botnet detection methods.* Computers & Security, 2014. 45: p. 100-123.