EMITTER International Journal of Engineering Technology
Vol. 7, No. 1, June 2019
ISSN: 2443-1168

Spatio Temporal with Scalable Automatic Bisecting-Kmeans for Network Security Analysis in Matagaruda Project

Masfu Hisyam, Ali Ridho Barakbah, Iwan Syarif, Ferry Astika S.

Department of Information and Computer Engineering,
Politeknik Elektronika Negeri Surabaya
Jalan Raya ITS, Sukolilo 60111, Indonesia
E-mail: masfu@pasca.student.pens.ac.id, {ridho, iwanarif, ferryas}@pens.ac.id

Abstract

Internet attacks are a frequent occurrence and the incidence is always increasing every year, therefore Matagaruda project is built to monitor and analyze internet attacks using IDS (Intrusion Detection System). Unfortunately, the Matagaruda project has lacked in the absence of trend analysis and spatiotemporal analysis. It causes difficulties to get information about the usual seasonal attacks, then which sector is the most attacked and also the country or territory where the internet attack originated. Due to the number of unknown clusters, this paper proposes a new method of automatic bisecting K-means with the average of SSE is 93 percents better than K-means and bisecting K-means. The usage of big spark data is highly scalable for processing massive data attack.

Keywords: IDS, matagruda, spatio temporal, automatic bisecting k-means, big data, spark.

1. INTRODUCTION

The internet is the greatest invention of this century. The massive development and the usage of the internet have affected in almost all aspect of modern life today. Internet connects many devices across the world. The rising number of internet users attract cyber-related crime such as information stealing, denial-of-service (DoS) attack, trojan, and malware. It is caused by the internet is not only used for information transmission but also for financial transcation. One of the popular ways to protect the transmission from those threats is intrusion detection system (IDS). Intrusion detection systems are software application to monitoring, analyzing, and protecting the network from the intruder or detect malicios behaviors [1] Matagaruda is a web-based application framework of intrusion detection system. Matagaruda provides four main features such as seeing or monitoring, learning internal traffic and generating internal rules, adapting the application framework tools and reducing misclassify alarm by using intelligence system [2].

However, intrusion detection system produces a huge amount of data and it becomes problems where traditional computing technology cannot process a large amount of data. While the traditional intrusion detection system still required monitoring tools but to process a large amount of data needs advanced technology called big data. Big data is a term for 3Vs defined: volume, velocity, and variety [3]. Volume means the amount of data. It is main problem to process using traditional computing technology. Velocity means the data rate to be processed The traditional computing is very difficult to handle fast data. Variety means the complexity of the data. The traditional computing is very difficult to process high dimensional of data from many sources. The data source are stored as log file or database format and need a various method for further analysis and visualization. One popular information model for cyber security is trend and pattern analysis. Using temporal, gives analysis of event distribution over time. And also pattern of distribution.

Matagaruda uses IDS (intrusion detection system) as the main component to monitor internet communication traffic. IDS is an application to analyze packet data through in a network. In other words, IDS is a sensor to detect anomaly activity or unusual activity that usually associated with the existence of cyber attack. When IDS detects a suspicious packet, it will generate a data called event. Event data consist of several types of attributes ranging from IP (internet protocol), time, severity or scale of attack vulnerability. Event data is usually stored in the form of log files or databases. Those event data need to be analyzed and visualized in further so that the data can be used for handling the cyber attack in the future. One useful method is trend and pattern analysis. In trend analysis, temporal spatial dimension is used for a time period, duration of time of the attack, and location information of event analysis. The IDS installed on matagaruda consists of 12 sensors and generate 500,000 to 4,000,000 data of event data every day. Then the method of large data processing will be difficult to do in a conventional way, therefore big data technology is needed to process and analyze the large data. Memory based computing [4] is a major component for processing in the large one which uses parallel computation to cluster the event data. In the results of clustering, we can get deeper insight and information from the cluster where the attack originated, then the types of attacks from each cluster. The commonly algorithm used is k-means as the clustering method. It is because the algorithm is very efficient. On the other hand, the results obtained are not so good. Hierarchical agglomerative clustering method is also commonly used for clustering to get good cluster results, but it has very slow processing speed [5]. Therefore big data with memory based to compute clustering is used to increase the scalability of data processing. We use data in 2016 for this experiment because BSSN formerly ID-SIRTI can not give the data, the newest data is very confidential.

2. RELATED WORKS

Many researchers conducted the study and proposed various methods of trend analysis and pattern to analyze network security data that have been done by researchers from various universities as follows:

The research was conducted by Shimeal and Phil Williams team from Carnegie Mellon University [6]. These researchers define information security models for analysis trends into five parts:

1. internal and external patterns are proposed to know the motive and purpose of the attacks that occur repeatedly. The internal pattern is used to find out the operand mode while the external pattern shows the issue of cybercrime on a larger scale.
2. temporal trend is used to determine the frequency distribution of attacks, types of seasonal attacks and the intensity of the ongoing attack activity. The processing technique uses cluster based on time dimension.
3. spatial trend is used to analyze the relationship between network attacks and physical locations where the attacker is located. It is very useful to know the types of attackers based on the location and the attack techniques used. Sometimes cyber war between countries can also be analyzed using spatial trends.
4. associational trends is used to analyze the relationship between one attack and another attack, such as the attack method used until the attack time.
5. compound trend is used to analyze the patterns of highly complex attacks consisting of broad dimensions such as space and time.

Disadvantages of the trend analysis is a report of the very large volume of data which needs big data to handle it.

The study was conducted by Zesheng Chen in 2015 [7]. the researcher saw network security as an important job in network management. One cyber security threat is the spread of malware. Their research focused on the model of malware deployment using network topology scanning techniques. It used graph analysis to visualize malware deployment and utilized spatial-temporal random processes to describe statistics of malware deployment on a network. The independent model can be used for temporal analysis while the Markov model is used for spatial analysis. Based on the comparison of calculations between the theory and simulation, the results obtained better accuracy.

The research was conducted by Guofei Jiang in Masschusetts[8]. Revealed that the event generated computer network is very large. While the computer network is a dynamic activity and evidence of the occurrence of distributed network attacks on some events or an event. The challenge is to correlate these events into a spatial-temporal data so that it can detect various scenarios of attacks that occur. The method used is Process query method used to correlate the distributed event data.

The research was conducted by Ferry Saputra [4], they designed big data architecture with BRO IDS (intrusion detection system) to capture data, while memory based processor is used for parallel processing. Several methods used such as k-means, k-means++ and bisecting k-means were compared to obtain botnet data clusters. Test results on the number of clusters ranging from single cluster have a speed of 60 seconds while 3 nodes have a speed of 33.7 seconds. the fastest clustering method is k-means and k-means ++ with 40 seconds of computation time while bisecting k-means needs 180 seconds. The accuracy of k-means, k-means ++ and bisecting k-means is obtained 97.7, 97.7 and 99.6. It concludes that bisecting k-means gives better accuracy result than the two previous methods.

3. ORIGINALITY

Analysis of network attacks is necessary for further action of an incident of cyber attacks, but that happens often the data about the attack is difficult to be analyzed and taken a conclusion, therefore to overcome the required analysis of trends and patterns on repetitive and seasonal attacks or when happened cyber war between countries [6], then we can analyze from which sector the attack appears and where to go. In addition, analysis for the attacked mechanisms occurring or sequences such as the source of attacks which are usually attacking to government sites or the financial sites when it is prior to trend and pattern analysis, therefore we offer a spatio temporal method for further analysis of cyber attack data. Because the cyber attack data is very large so it is difficult to be processed using the traditional method and then we offer a scalable automatic bisecting-kmeans method capable of processing large data and determining the number of data clusters automatically. The scalable automatic bisecting-k means method consists of bisecting-kmeans and moving variance analysis. the bisecting-kmeans method itself is a very effective divise hierarchical clustering method for processing large data rather than regular agglomerative hierarchical clustering [9]. The method of bisecting-kmeans processing the data set formed from a large cluster and then separated into several clusters accordingly with a certain iteration limit [10]. In addition, this method can determine the number of clusters that reach global optimum by using evaluation of moving variance or known as automatic clustering [11]. But because the process of clustering requires a large computing resource, the role of big data can help speed up computing by using multiple computers in parallel.

4. SYSTEM DESIGN

The system design in this study is divided into 3 main processes, there are Data Collections, Transformation, Data processing and Visualization as illustrated in the system design in Figure 1.

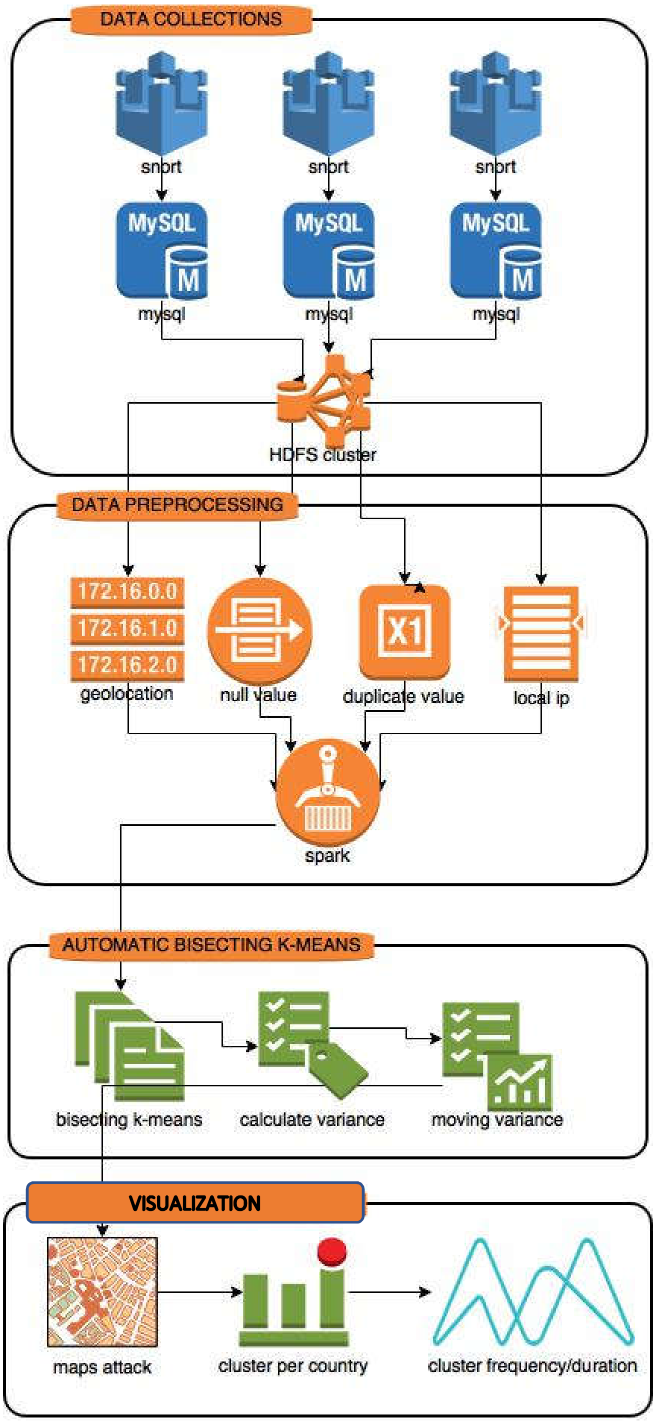

Figure 1. Block Diagram of Spatio-Temporal Network Security Analysis on Matagaruda

Analysis of network attacks is necessary for the further action of an incident of cyber attacks. But it is often happend while the data attack are difficult to be analyzed and taken a conclusion. Therefore to overcome the required analysis of trends and patterns on repetitive and seasonal attacks or when happened cyber war between countries [6], then we can analyze from which sector the attack appears and decide the next step.

In addition, analysis for the attack mechanisms or attack sequences such as source of attacks which is usually attacking to government sites or the financial sector that Prior to trend and pattern analysis, therefore we offer a spatio temporal method for further analysis of cyber attack data. Because the cyber attack data is very large, it is difficult to be processed using the traditional way. Then this paper offers a scalable automatic bisecting-kmeans method, which is capable of processing large data and determining the number of data clusters automatically. The scalable automatic bisecting-k means method consists of bisecting-kmeans and moving variance analysis. The bisecting-kmeans method itself is a very effective divisive hierarchical clustering method for processing large data rather than regular agglomerative hierarchical clustering [9]. The bisecting-kmeans method process the data which is set formed from a large cluster and then separated it into several clusters accordingly with a certain iteration limit [10]. In addition, this method can determine the number of clusters that reach global optimum by using evaluation of moving variance or known as automatic clustering [11]. The clustering process requires a large of computing resource, therefore the role of big data can help speed up computing by using multiple computers in parallel.

4.1. Data Collection

Matagaruda currently uses data warehouse technology to process data from 12 sensors installed in NAP (network access provider) and ISP (internet service provider). This research will contribute big data process in case of Hadoop cluster. After data from the data warehouse are moved in batch mode or offline processing, the data will be sampled by using spark for data clustering or grouping based on attacker IP addresses, attacker's purpose, attack type, port number, services, protocol, geolocation and attack time. The clustering method is used the k-means bisecting method (a group of hierarchical clustering algorithms divisive), which processed data set formed from large clusters, then separated into several clusters in accordance with certain iteration limits. This method can determine the number of clusters that reach global optimum by using evaluation of moving variance or known as automatic clustering. The role of big data can help speed up computing by using multiple computers in parallel because the process of clustering requires a large computing resource.

4.1.1. Matagaruda

Matagaruda as a framework for network security analysis. Matagaruda consists of several analytical modules such as statistics for attacks from IDS. Intrusion Detection System (IDS) is a software or hardware application that can detect suspicious activity in a system or network. IDS can inspect inbound and outbound traffic in a system or network, perform analysis, and look for evidence of intrusion experiments (intrusions) [12].

4.1.2. Sensors

The sensor consists of an IDS snort that detects attacks from the entry internet, in this case, the NAP (network access point) / ISP (internet service provider). SNORT is one example of software from NIDS that matching packets to the rule and determining a packet or the coming attack into the network as an intrusion or not. SNORT is able to analyze real-time traffic and packet logging on the network, capable of performing protocol analysis, matching content, and can also be used to detect various types of attacks and checks such as buffer overflows, stealth port scans, CGI attacks, SMB \Probes, OS Fingerprinting attemps, and much more related to network attack experiments [13].

4.1.3. Database

The database is used for storing the data from the IDS snort sensor to the warehouse database which is using MySQL database.

4.1.4. Hadoop

In this paper, Hadoop is responsible for storing data. Hadoop has the component such as HDFS. HDFS is a distributed file system which can handle a very large data storage. Hadoop becomes the place of all data so that it can be analyzed by various tools for various purposes in order to get a detailed result and meet the needs of the user [14].

4.2. Data Preprocessing

Procedure data is divided into several stages such as deleting null value data, duplicate value, local IP. After the cleaning phase has been completed, then the data transformation stage is done by grouping data based on ip_src and ip_dst and also calculated frequency and duration for each attack stage. The last stage is to look for geolocation from IP source, IP destination, city, and country. All the preprocessing used Spark, which is the next generation of MapReduce. MapReduce is very effective for summarizing queries and analyzing large amounts of structured data. Hadoop calculations using MapReduce is slower than In Memory Computing, where spark enters. Spark was developed at UC Berkeley AMPLab in 2009 and sourced in 2010 [15], Apache Spark is a powerful Hadoop data processing engine designed to handle both Batch and streaming workloads in no time. In fact, in Apache Hadoop 2.0, Apache Spark runs the program 100 times faster in memory and 10 times faster on disk than MapReduce [16].

4.2.1. IP sector dataset

Sector IP data is a list of companies, institutions, or governmental institutions with domain names, and ip addresses. Here is a list of sectors in table 1 monitored by the sensor from matagaruda :

Table 1. Lists the sectors monitored by Matagaruda

Number	Sector
1.	defense
2.	energy
3.	finances
4.	health
5.	information
6.	governments
7.	transportation
8.	food

4.2.2. Dataset event

The data generated from the IDS snort is usually a log file or database if configured with MySQL as storage. In this study the data generated event will be accommodated by MySQL database with data that has been summarised with data protocol and services following attributes of the dataset event.

Table 2. Snort event table scheme

1	long_date	string
2	code	string
3	cid	int
4	sid	int
5	signature	int
6	ip_src	bigint
7	ip_dst	bigint
8	tcp_sport	int
9	tcp_dport	int
10	udp_sport	int
11	udp_dport	int
12	signature_name	string
13	signature_priority	int
14	protocol	int
15	detail	boolean

The 15 attributes to be used for spatio temporal analysis is a long_date attribute, ip_src, ip_dst, signature_name.

4.2.3. Data Processing

Procedure data is divided into several stages such as deleting data that is null value, duplicate value, local ip. After the cleaning phase is completed, it is followed by transformation data stages to grouping data based on ip_src and ip_dst and also calculated frequency and duration for each attack stage. The last stage is to look for geolocation from ip source, ip destination, city, and country.

4.3. Automatic Bisecting K-means

Clustering is divided into several stages as illustrated in Figure 1 with the main process of Automatic Bisecting K-Means is run on top of the Spark by using Yarn on Spark or job scheduler, and some methods are run parallelly by using in memory computing models such as bisecting k-means and moving variance. The following sections on the data processing :

4.3.1. Bisecting k-means

Bisecting k-means is a hierarchical clustering method commonly used to process large data [5], it is possible because bisecting k-means is a type of divisive algorithm and can be limited in a number of iterations [10]. In the bisect process or divide a subset into two clusters (C_1 and C_2) it is used k-means algorithm with initial centroid fixed and paired, the initialization of the centroid is with upper bound and lower bound if d as dataset:

$$C_1 = \min(d) \tag{1}$$

$$C_2 = \max(d) \tag{2}$$

Where C_1 is the first centroid and C_2 is the second centroid. Centroid value is fixed for every k-means process, the next process is error calculation with SSE (sum square error) :

$$\sum_{j=1}^{K} \sum_{x_i \in C_j} || x_i \quad \mu_j||^2 \tag{3}$$

where μ_j is the centroid of C_j, Sum of Square errors is a metric used to measure clustering effectiveness. Where K is the number of observations and x_i is the value of the i-th observation.

4.3.2. Moving variance

Moving variance is one of the automatic clustering methods that can achieve global optimum but this calculation must be paid handsomely with big computational cost [11]. So to overcome this problem is used parallel processing techniques using spark as framework. There are two stages to find the optimum K value. The optimum K value is calculated by the variance of each (V = 1 ... n) where n is the maximum number of iterations or altitudes of a hierarchical clustering tree. Variance is divided into two, there are variance within cluster and variance between cluster :

$$V_c^2 = \frac{1}{n_c \quad 1}\sum_{i=1}^{n_c}(d_i \quad \bar{d}_i) \tag{4}$$

v 2 = variance in cluster c c
c = 1..k, where k = number of clusters
nc = amount of data on the cluster c
d_i = ith data on a cluster
$\bar{d}_i$= the average of the data in a cluster

while for variance within cluster denoted by:

$$V_w = \frac{1}{N \quad k}\sum_{i=1}^{k}(n_i \quad 1).V_i^2 \tag{5}$$

for v_w = variance within cluster N = sum of all data can be denoted by the above formula while searching for variance between cluster :

$$V_b = \frac{1}{k \quad 1}\sum_{i=1}^{k}n_i(\bar{d}_i \quad \bar{d})^2 \tag{6}$$

with d = the average of in, to find the variance of all clusters it is denoted by

$$V = \frac{V_w}{V_b} \tag{7}$$

4.3.3. Valley Tracing

The last stage of automatic clustering is determining the optimum number K of variance [11], while the method used to determine global optimum is valley tracing :

$$(v_{i-1} \geq v_i) \cap (v_{i+1} > \; v_{1)} \tag{8}$$

with i = 1 ... n where n is the number of stages.

4.4. Visualization

The event data generated by IDS has a great number, therefore, visualization becomes a challenge. Some visualization techniques commonly used for network security attacks are using maps with the attacker plotted and filtered by time. In this research, the attacker's data will be displayed based on geolocation, and it will also be grouped according to the clustering result from previous process.

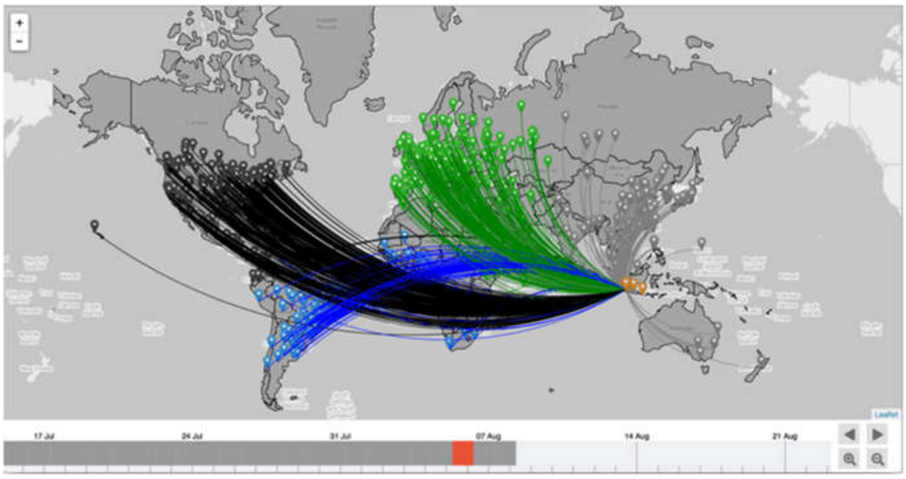

Figure 2. Map Analysis Using Spatio Temporal

If the slider at the bottom is shifted then the displayed data will change according to the selected date so that the spatio-temporal analysis is easier to do.

5. EXPERIMENT AND ANALYSIS

In this paper, the experiment was conducted by taking samples of data from several dates. This is done to facilitate the spatio-temporal analysis of network attack data. Before the spatio-temporal experiment is done, the trends and patterns of the internet sequence from entire data have been observed, the following day-to-day trend charts are shown below.

5.1. Top 20 Alarming Signatures

Agregation metric can be used to visualize the traffic attack pattern, identify DoS attacker, shows the misconfigurations, and shows signature name from data attack which is processed by sensors. Twenty top signatures can give information about the condition of network, then tuning IDS sensor become possible.

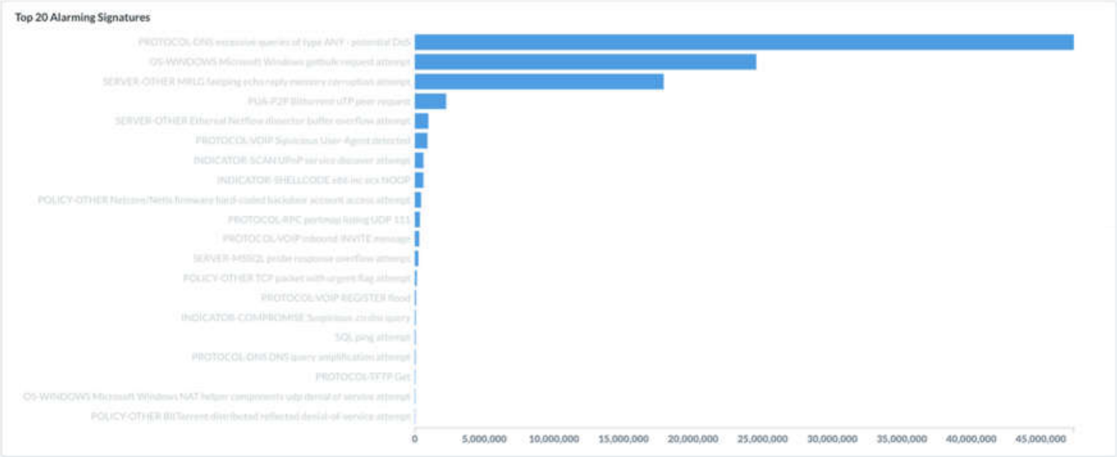

Figure 3. Top 20 Alarming Signatures

From Figure 3, the type of signature names are mostly happened in *PROTOCOL DNS excessive queries of type ANY – Potential DoS* which has cases almost a half or 45 million from 97 million of total cases, then followed by *OS-WINDOWS Microsoft Windows getbulk request attempt* which has 25 million attacks. Therefore it can be conclude that mostly attack happen based on DOS system.

5.2. Top 20 Alerts by Date Metrik

Agregation metric can be used to give the pattern and identify the harm activity on weekend days or in holiday season. The high number of errors in authentication at the beginning of weekdays can be categorized as normal. However, the number of authentication errors on weekend days are interesting to observe. Figure 4 shows the number of attacks happened in every dates.

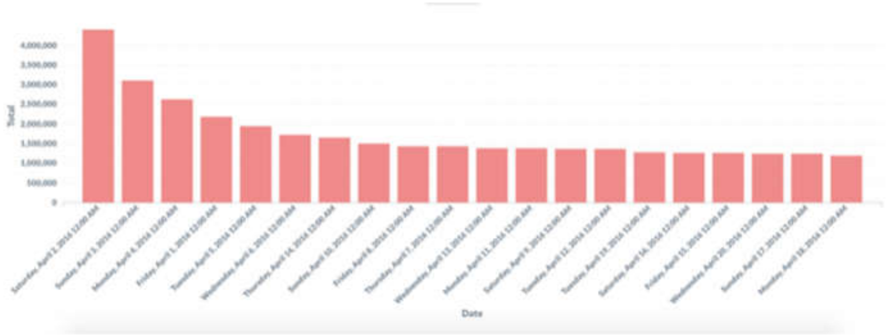

Figure 4. Top 20 Alerts by Date Metrik

Based on Figure 4, it is shown that the number of attacks have domination on weekend days from Saturday until Sunday. However, it is happened until Monday. The attacks happen from Tuesday until Friday are relatively low.

5.3. Alerts by Source IP

Building dataset from source ip can be used to identify reconnaissance or sniffing based on the time. Moreover, it also can be used to identify DoS attacks, therefore, this bias matric is processed by tunning IDS sensor and eliminated false positive.

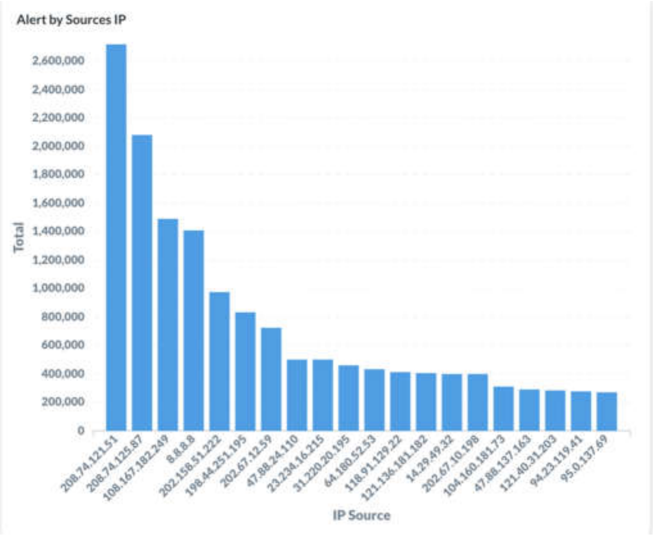

Figure 5. Alerts by Source IP

From Figure 5, is shown that IP 208.74.121.51 has been done the most attacks with number of 2.6 million attacks. This indicates DoS attack. Furthermore, the ip 208.75.125.87 which has 2.1 million of attacks. The top two attacks are from same block IP which indicate the attacks are from same person but using different host.

5.4. Alerts by Destination IP

The data of attack's purpose can be used to get information which the host has a lack and become target from attacker around the world. This system has function re-checking the attack's destination host like port blocking or adding the firewall. It also can be used to give warning for the network server administrator management.

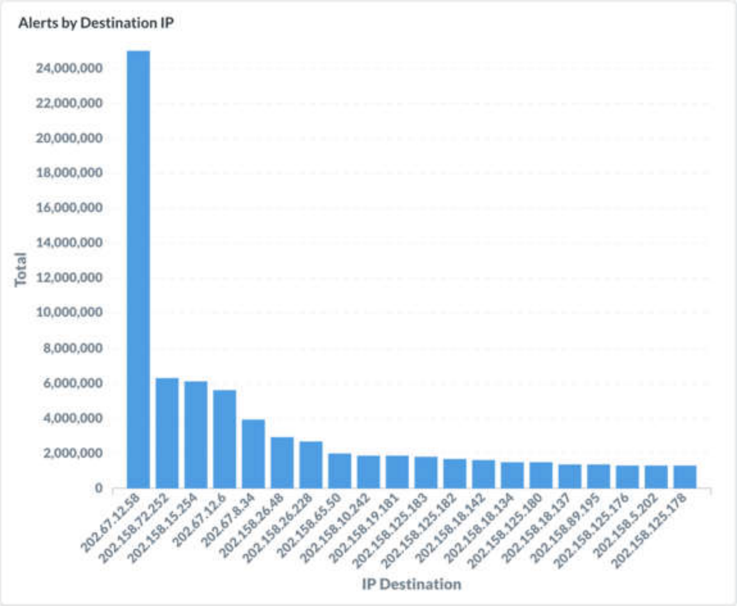

Figure 6. Alerts by Source IP

From Figure 6, it shows that ip 202.67.12.58 has 24 million of attacks and becomes the most attacks IP. This indicates DoS attack. It is followed by IP 202.158.72.252 which has 6 million of attacks. Based on the top two IP which came from the same block IP, it indicates the attacks are from the same person but using different host.

5.5. Source by IP Country

The geolocation data can be a strong indicator when it is used together with the others metric data. By utilizing the geolocation data, the system can get information of the attacker's country and city, even from ISP and AS. The visualization of spatio temporal from geolocation data is shown in Figure 7.

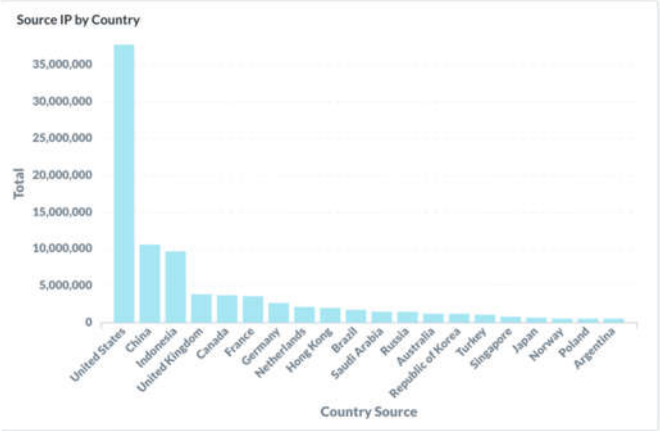

Figure 7. Source by IP Country

Based on Figure 7, it shows that the country with the most number of attack is United States, following with China and Indonesia. The number of attack from

US is almost 39 million cases, from China is 10 million of cases and 9 million of cases from Indonesia.

5.6. Total Number Alarm by Hour/Day

Not only the number of attack every day, but the system also gives information about the number of attack per hour. Aside from the number of attacks per day the system can also see the number of attacks per hour. these data can give information the patterns of activity per hour of the attack. Therefore, our system are more vigilant at certain hours. Ordinary statistical aggregation can be used to find out IDS sensors system work. The statistical aggregation can also be used to find out the abnormal activities that which indicates a masiv attack at certain times related to other parameters.

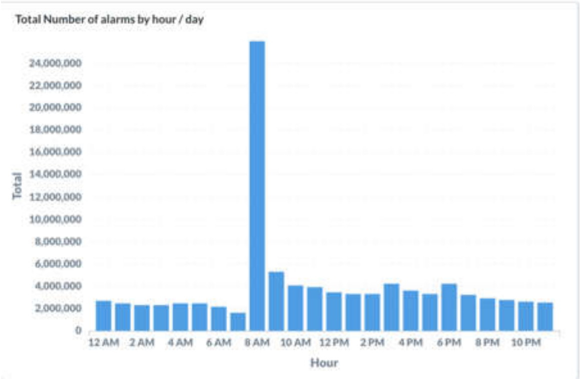

Figure 8. Total Number Alarm by Hour/Day

From aggregation results per hour/day, it can be concluded that the attacks are usually occurred at 8 o'clock with the intensity of the attacks of 24 million cases, then decreased quite dramatically at the rate of 4-6 million attacks until 10 o'clock at night, and it continues to decline until 7 am.

5.7. Alert Categorized by Severity

The attacks are categorized based on the risk level (severity) with the intension handling system priority with the intention of handling a high risk attack.

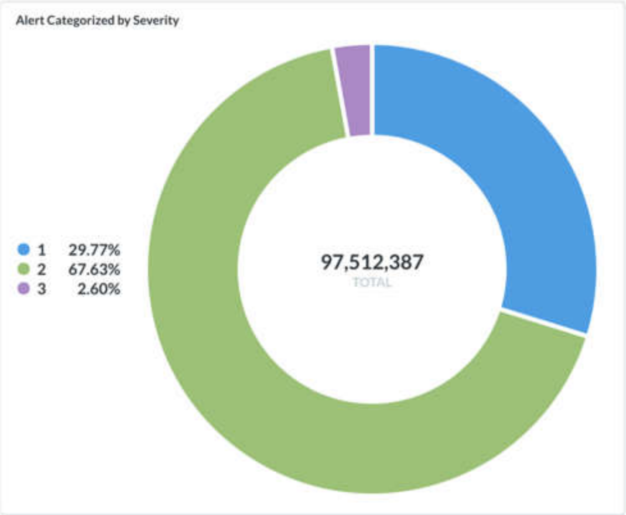

Figure 9. Alert Categorized by Severity

From Figure 15 it can be seen that the level of risk of the medium or green has mostly happened, then the second is at the risk of low-level attack, and the smallest is an attack with the highest risk or high severity.

5.8. Attacks per Sector

If event data is integrated with sector ip data then we can observe which sector gets the most attacks. The following graphs of attacks per sector:

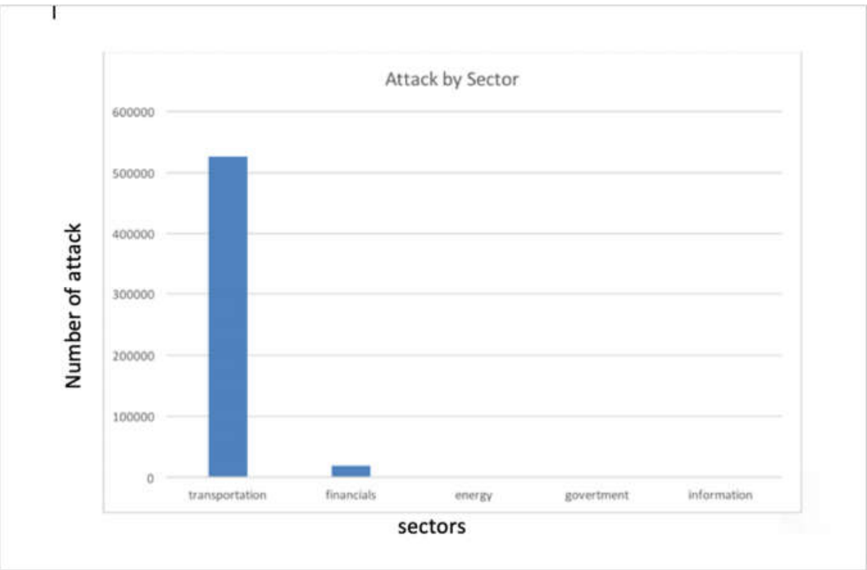

Figure 10. Sector-based attacks

From Figure 10 it can be seen that the transport sector that has the most attacks with a total of 526254 attacks, followed by the financial sector that got 18957 total attacks.

5.9. Spatio-Temporal Analysis

Spatial-temporal analysis of all data collected for 3 months and it will produce a very large analysis, therefore we will only take samples 3 days from 05-08-2016, 06-08-2016, 07 -08-2016. Here's the visualization of spatio-temporal:

5.9.1. Map of attack

The visualization is not only displaying the attack map, but also displaying a clustered attack automatically using the automatic bisecting k-means method. From the clustered data, they can be grouped into color. Here is the attack map on 05-08-2016 :

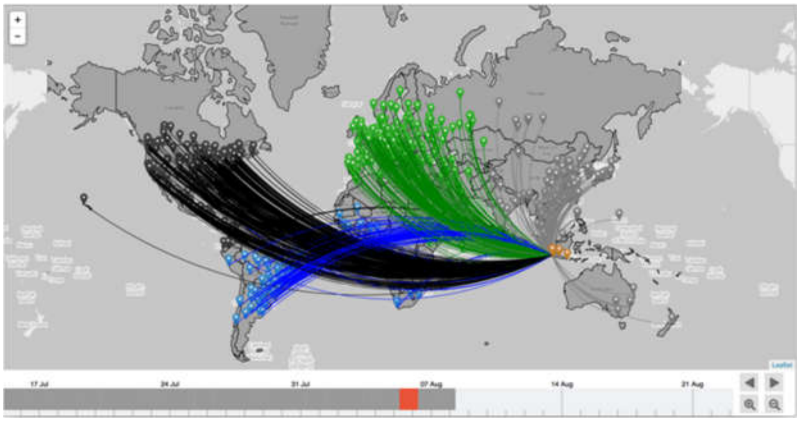

Figure 11. Map of cluster attacks on 05-08-2016

Figure 11 appears that the number of clusters are 4 pieces of cluster, and each cluster is clustered by region region. The number of clusters for each day is clearly different because the attacks are also different, following the attack map on 06-08-2016:

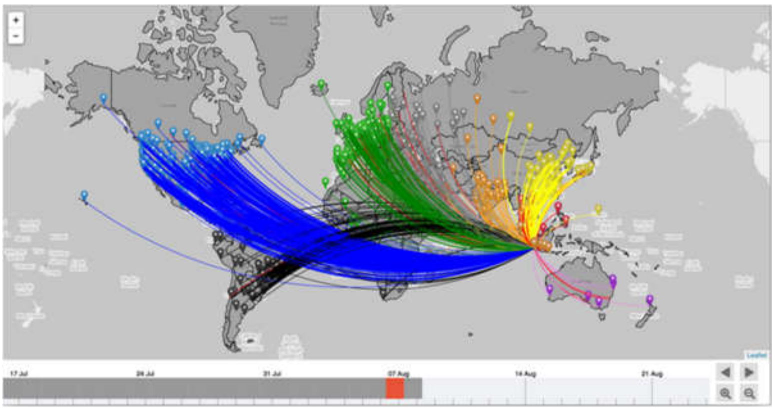

Figure 12. Map of clustered attacks on 07-08-2016

Figure 12 appears that the number of clusters is 8 clusters and each cluster is clustered by region region. Also seen attacks dominated from the North American territory as in the blue data. following the attack map on 07-08-2016:

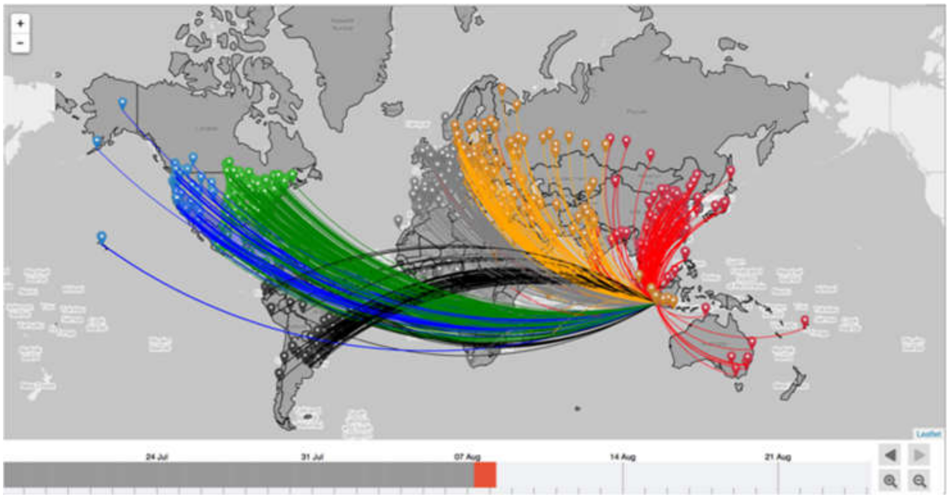

Figure 13. Map of cluster attacks on 08-08-2016

From Figure 13 appears that the number of clusters is 6 clusters and each cluster is clustered by region region. Also seen attacks dominated from the western region of europe and eastern europe as in the data are gray and orange color, and also attack from east asia region like from china as shown with red color.

5.10. Performance Analysis

Performance analisys consists of two main parts which first automatic bisecting k-means accuraccy and the second one is time computation. In each experiment of the clustering method, the dataset is divided by date with the assumption for k-means and bisecting k-means method that the number of k is determined by 6 clusters, while for automatic bisecting k-means the number k will be automatically obtained. Based on experiments for 3 different methods of bisecting k-means, k-means and automatic bisecting k-means. only k-means uses the initial centroid randomly. While k-means k-means and automatic k-means uses the initial centroid with minimum value and maximum value, then the result of k-means experiments will always change, while automatic bisecting k-means will always produce a constant initial centroid. It is necessary to create hierarchical clustering.

Clustering validation is essential for performance analysis of a clustering method and the usefulness of cluster validation is to avoid looking for patterns in noise, comparing clustering algorithms, and comparing two clusters of several methods to measure the accuracy of a clustering method one of which is by sse or sum squre of error. SSE is good for comparing two clusters or two clusters. The experimental results in Figure 14.

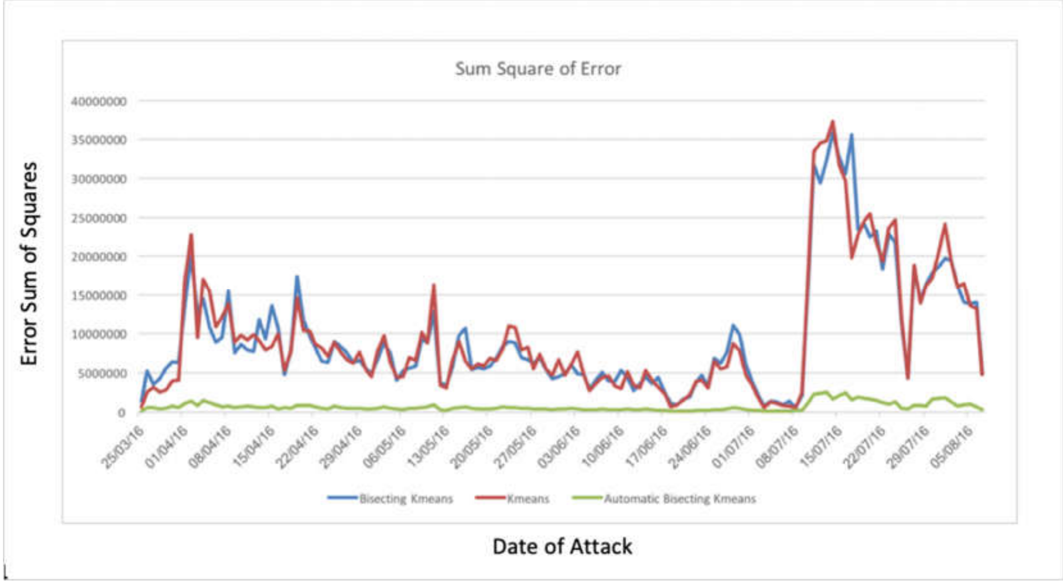

Figure 14. Sum Square Error Validation

In Figure 14 it is seen that bisecting k-means obtains a high SSE value in which the result of clustering data has not been well separated when compared to k-means which have a smaller SSE value. By using automatic bisecting k-means, the SSE value becomes smallest than bisecting k-means and k-means.

Table 3. Statistics Sum Square of Error

	bisecting k-means	k-means	automatic bisecting k-means
min	595.346	511.466	81.798
max	35.885.904	37.327.685	2.507.915
average	9.532.910	9.548.692	645.332
median	6.774.990	7.038.094	479.778

From Table 3 it can be seen that from the three methods, the smallest average of SSE is automatic bisecting k-means and the least of median value is the automatic k-means. But of all SSE value method is still classified as large classes because the dataset is not separated, therefore from it need to do outlier detection and outlier remover.

The next experiment is a computation time performance test. In this experiment, the time required to complete the calculation of each method will be compared. This experiment is very useful to know the computational speed of each method therefore the scalability of the clustering method used can be measured.

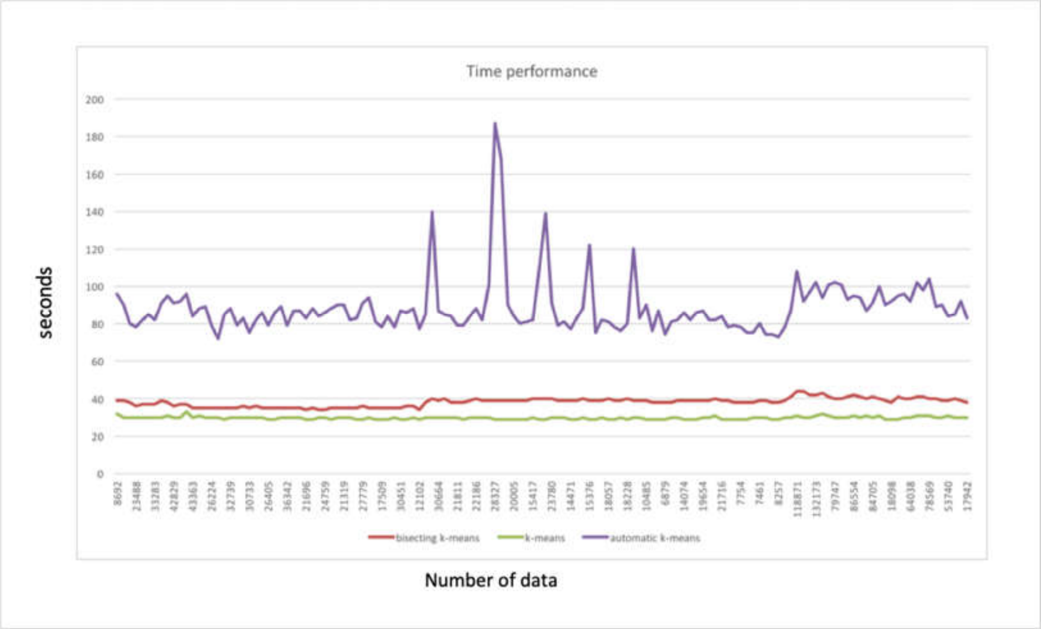

Figure 15. Time Performance

In Figure 15 it is seen that k-mean is the fastest method of all methods used, then it is followed by k-means bisecting method where its performance approximates k-means method, while the slowest of the three methods used is automatic bisecting k-means. This is very reasonable because it has to compute bisecting k-means, calculate variance, and calculate valley tracing to determine the optimal k value. Then from the k-means automatic bisecting experiment, the results obtained are the k values of each experiment by the date, therefore that each date gives different k values depending on the pattern of attack data that occurs on a daily basis.

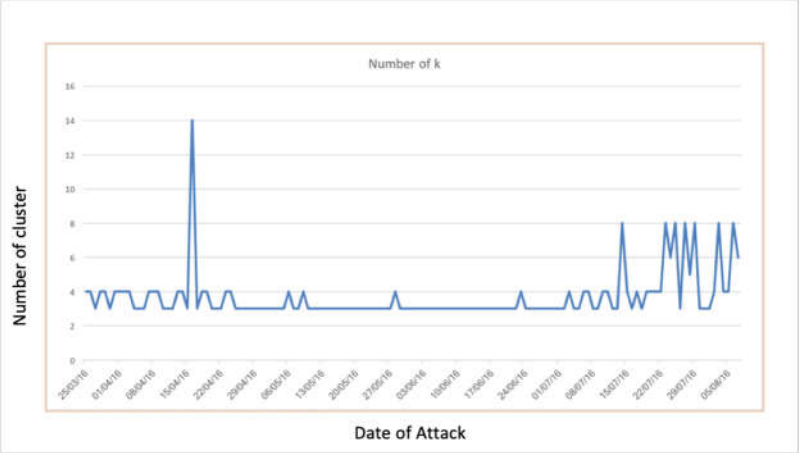

Figure 16. Number K from Automatic Bisecting K-Means

In Figure 16, it is seen in April the total value of k reaches the highest value that is 16, then stabilizes at the number k 3 to 4 in the following months. The value increases again to an average of 8 in July and August.

6. CONCLUSION

This study has collected and analyzed the spatio temporal data from matagaruda project with the aims to monitor network attacks occurred in Indonesia and the trend of internet attacks that occurred in Indonesia can be analyzed. The most widely targeted sector analysis of internet attacks in Indonesia also can be seen and analyzed. The most crucial is the spatio temporal analysis used to find out the attacks and where the attack originated. Spatio temporal analysis using automatic bisecting k-means gives the number of cluster attacks automatically based on data variance, therefore it can provide more accurate information with SSE value 93 percents is better than k-means and bisecting K-means. The value of K (cluster) in automatic bisecting is generated automatically, it depends varians value. While value of K (cluster) in k-means and bisecting K-means is 6 clusters. Automatic bisecting K-means also can be used to calculate data with increased scalability as needed. K-means automatic bisecting method takes a relatively longer time than k-means and bisecting k-means with an average of 3 times longer than k-means and bisecting k-means.

Acknowledgements

This research is provided by ID-SIRTI or Indonesia security responses team.

REFERENCES

[1] Zarrabi and A. Zarrabi, **Internet Intrusion Detection System Service in a Cloud**, *IJCSI International Journal of Computer Science,* vol. 9, no. 5, p. 1, 9 2012.

[2] F. Astika, I. Winarno and M. B. Muliawan, **Implementing Network Situational Awareness in Matagaruda**, in *International Electronics Symposium (IES)*, Surabaya, 2015.

[3] R. Zuech, T. M. Khoshgoftaar and R. Wald, **Intrusion detection and Big Heterogeneous Data: a Survey**, *SpringerOpen Jurnal,* vol. 2, no. 3, p. 4, 2015

[4] F. A. Saputra and A. Abdillah, **Big Data Analysis Architecture for Multi IDS Sensors using Memory based Processor**, Surabaya, 2017.

[5] M. Steinbach, G. Karypis and V. Kumar, **A Comparison of Document Clustering Techniques**, Minnesota, 2000.

[6] T. Shimeall and W. Phil, **Models of Information Security Trend Analysis**, Piitsburgh.

[7] Z. Chen and C. Ji, **Spatial-temporal modeling of malware propagation in networks**, in *IEEE Transactions on Neural Networks*, Atlanta, 2005.

[8] G. Jiang and G. Cybenko, **Temporal and spatial distributed event correlation for network security**, in *American Control Conference*, Boston, 2004.

[9] S. Harifi, **Comparative Study of Apache Spark MLlib Clustering Algorithms**, in *Data Mining and Big Data: Second International Conference, Fukuoka, 2017.*

[10] Y. Zhuang, Y. Mao and C. Xin, **A Limited Iteration Bisecting K-means for Fast Clustering Large Dataset**, Texas, 2016.

[11] A. R. Barakbah and K. Arai, **Determining Constrains of Moving Variance to Find Global Optimum and Make Automatic Clustering**, Surabaya, 2004.

[12] M. Tiwari and A. Bharti, **INTRUSION DETECTION SYSTEM**, in *International Journal of Technical Research and Applications*, New Delhi, 2017.

[13] S. Chakrabarti, I. Mukhopadhyay and M. Chakraborty, **Study of snort-based IDS**, Mumbai, 2010.

[14] T. Qureshi, **Big Data and Hadoop**, in *International Journal of Computer Application & Applied Sciences*, CollegeFaisalabad, 2015.

[15] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker and I. Stoica, **Spark: Cluster Computing with Working Sets,** 2010. [Online]. Available: http://people.csail.mit.edu/matei/papers/2010/hotcloud_spark.pdf. [Accessed 2018].

[16] Apache, "**Apache Spark**," Apache, [Online]. Available: https://spark.apache.org. [Accessed 28 04 2018].