International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Advanced Network Security Analysis (ANSA) in Big Data Technology

Shivi Sharma, Ashish Sharma, Hemraj Saini

Abstract*: Big Data has caught the attention of research, science, and business world due to the advancement in digitalization. With the evolution of the Internet of Things (IoT), data is increasing by massive amounts every day. In the big data environment, securing a large amount of data has become a challenging issue in both security and research industry. In this paper, a framework has been proposed to inspect malignant information and suspicious activities traveling over the networks by utilizing Hive Queries. This framework's procedure loads activity information into Hadoop Distributed File System (HDFS) through a Hive database thus examining the information. This information is sorted as IP Wise, Port Wise, and Protocol Wise. Hive queries will help to achieve these three goals:- 1) Traffic classification 2) Interrupt Identification 3) analyzing of network traffic. Using this framework provides users' a benefit of being able to investigate Big Data and helps them to detect attacks. Therefore, this framework will allow prevention of network attacks and enable real-time detection in a Big Data environment.

***Keywords* : Big Data, Hadoop, Network attack, Network Security Analysis**

I. INTRODUCTION

With the requirements of the modern day technology, the need for Big Data has become quintessential for its operation. As the volume of information is increasing every day, the threat to its security is also increasing [1]. Since existent traditional systems are unable to assess and inspect such large collections of dataset, this task has become a huge challenge for the network security [2]. Data security is one of the topmost priorities of network services, to ensure that the users have faith in their services when using to exchange personal information and various other transactions [3]. Big Data security analysis assesses the large volume of security data from an organizational perspective, which often requires expensive data and a is a powerful IT infrastructure [4]. There have been many researches regarding the same by scientists, researchers, programmers in order to build a sanctuary where the data is safe and services can be offered to users which provides trusted security to their information.

The fundamental objectives of a security network are to preserve and maintain Confidentiality, Integrity, and Availability (CIA) to the users [5].

The possibility of detecting attacks and violations in a network in a real time environment will enhance the capability of a user to process and transfer information and dataset. Real time solving efficiency will automatically bring down the threats and will instill transparency between the users and services [4]. Throughout in this research, an experiment is conducted to observe the traffic volume and analyze the network for any potential threats through Hadoop Distributed File System (HDFS) with the aid of Hive databases. HDFS is efficient in its ability to categorize and analyze their results due to the parallel computation that is carried out on the information available. This parallel operation helps in identifying any suspected violations in the network [11][12] .The power of computation can be further enhanced with more resources and powerful technology to the point where the security threats can be dealt/handled within a real-time interface. A structured query language called Hive Query Language simplifies the Map Reduce programming and extracts the metrics of the available dataset. [6][13]. In this paper, analysis of security network through HDFS environment is done to detect contents violating security network, cyber-threats, and other attacks to preserve the integrity and security of this huge big data network systems and services[5]. Following sections will tend to the details of this research paper's objective, proposed methodology to carry out the analysis and the observed results' assessment.

II. OBJECTIVES

The main aim of this study is to find a semantic gap in which modern networks work at the transport layer while the existing frameworks work at the network layer. To address this gap and empower forceful enhancements, here presenting the idea of bundle stream catch, in view of the key reflection of the bit stream and processing it by the Hadoop system. Following are the key goals :

1. **Traffic capture**: The volume of dataset is identified from the streams and is stored to be analyzed further for threats and other operations, while maintaining unique elements from the collection of information. It is further distributed and the traffic volume is characterized by a specific classification purpose.

2.**Intrusion identification**: The presence of anomalies in Big Data is a common sight with the ever-growing technology and increasing volumes of information. Intrusion identification plays a vital role in detecting these anomalies in the network by monitoring the traffic at packet level and comparing it with a pre-existing information of similar attacks.

Manuscript published on 30 August 2019.
*Correspondence Author(s)
Shivi Sharma, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, Solan-173234, Himachal Pradesh, INDIA,
Ashish Sharma, SAP SD Consultant , IBM Pune, Maharashtra -411006,
Hemraj Saini, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, Solan-173234, Himachal Pradesh, INDIA,

However, it encounters two major issues. It fails to detect slow attacks, avoiding the triggering of security alarms, or the limits set by the Intrusion Detection System (IDS). In addition, it does not recognize those threat packets, which share no similar characteristics with the database of known attacks.

In addition, with the increasing volumes of information with varied data operating at unpredictable velocities, it has become a concerning challenge to the security of the network. But with the help of the research and studies by Baijan Yang and Tongling Zhang, a model has been proposed to overcome these issues [6]. In this model we propose to integrate their method in identifying lethal intrusions and detect variety of threats to the Big Data network.

3. **Network Security Analysis**: The final objective is to investigate the network traffic, calculate its metrics, and categorize it according to the requirement of different ISPs.

However, the concerning rise of new variety of threats has issued a need of an improved processing system to analyze and assess the dataset. A combined framework has been proposed to process Big Data with a semantic connection and inference methods which can alter the way network analyzed and help overcoming various threats [9].

The security analysis of Big Data undergoes three operations. Firstly, the dataset is analyzed by the traditional security methods satisfying the performance requirements. Then, the existing analyses are implemented on the Big Data system, processing them into storage clouds. Finally, it is pre-processed with new models optimizing for the modern day developmental requirements.

The semantic distinction in the security analysis characterizes it into various semantic factors which may be based on concepts, mode of execution or on its degree of explicitness [7][14]. Semantic values help in recognizing a packet or a threat in a more distinct classification [8].

Map Reduce takes the data-set and divides it into unique elements which improves the efficiency and the scalability of HDFS, providing a powerful basis to do the analysis on the information which is stored in the clouds [4].

It is responsible for the computation and data processing of Hadoop. Fault tolerance is built in Map Reduce enabling it to run on commodity hardware. Tasks are also assigned to the nodes through this software by distributing data across various nodes ensuring that the results are observed and consolidated. [8] [5]

Map Reduce has five components [10]:

• **Name Node**: Indexing of data is done by name node similar to the server where all data is uploaded. It stores all the data in the HDFS environment and can track it across the packets.

• **Data Node**: Every actual data is stored in this node and it is well connected to the name node during the startup

• **Job Tracker**: It is built on a master assigning and executing operations. Tasks include splitting different jobs like IP count, calculation of data and port wise capture etc.

• **Task Tracker**: Job tracker assigns the job to task tracker.

• **Secondary Name Node**: It is used for data security purpose. It is also a replication of the name node; as if the name node fails, then the secondary name node is used.

III. RELATED WORK

Network security analysis is complex for users and professionals due to the generation of large size of data in Big Data systems. For the measurement and improvements in security, different security tools are used such as application logs, event logs, firewall logs, and other security logs that help in detection of anomalies and inconsistencies in data [3] [7]. These tools are effective and efficient in handling the collected data and its processing, which often need high resources of system and strong tools for analysis of the data.

On the other hand, existing systems and conventional tools are incapable of analyzing hand handling unstructured data. Large set of data requires a high level of security at both data layer, network layer, and transport layer [1].

The need of security in transport layers is crucial because of the threat of intrusion and attack from unauthorized access of users, non-users, and hackers. Once the dataset is protected on these layers, it can be said that the network level security is controlled and managed [9].

On the other hand, if the security is less effective on network layers, then the valuable data will become useless and an overhead of resources for other important datasets and applications [2].

Another issue that has been found after using these tools in security control is that untrained users are quite reluctant of data management, and for them using these tool may not be effective in Big Data systems and large datasets. Some studies have focused on the implementation of Intrusion detection system (IDS) in the Big Data systems for the protection of large datasets. The presence of anomalies in the technology and large volumes of data requires data protection and system security at both the client and server level [5]. The detection of threats and intrusions in the system will be useful for the users and data handling professionals to understand the point where the security might be low. The IDS is widely used in various businesses, and have many techniques to detect and control data loss and attacks which can affect the systems [7] [9].

A few research studies have focused on data traffic volume as well as its capturing large volumes of data is analyzed and stored for operations, along with handling threat and managing the information present in data. The data traffic also helps in determining trends through figures which can be used for security analysis and pertinent data management [4] [5].

A research study emphasized on network security analysis for Big Data systems by classifying various internet service providers and their relevance and roles. The increase in new and uncertain risks and threats may not be in the favor of Big Data systems, but in order to overcome the, contemporary network frameworks are required which build a connection between Big Data systems and networks enabling security to protect the systems[6].

A number of factors deal with network security analysis and identification of threats in transportation of datasets from one system to another. Data indexing is done to list the data where it is uploaded.

Data indexing is one of the most useful way to keep a track of data in Hadoop's environment and HDFS because through indexation of large volumes of data, the missing values are found easily via binary or linear searching and anomalies are easier to determine from the datasets [10][11].

Big Data security is essential because of the high risk of exposure of large volumes of data and revelation of confidential information of millions of users. The Big Data systems are present in mostly all big enterprises where patterns and trends are identified without any human interaction, and are based solely on the analysis through computation [8]. The big data systems are adopted to cater massive amount of data. Hadoop is one of the most effectively used tool for big data analysis and it helps in disclosing important features and aspects associated with interaction and behavior of human beings [9]. Therefore, the adaptation of security should be ensured in the organization along with planning to take strong security measures that will prevent the leakage and misuse of data from the systems [10].

Within Big Data systems, there are some useful attributes of security such as integrity of data, availability, accessibility, and most importantly confidentiality. The breach of confidentiality and any other attribute will be considered as exploitation and loss of security [10] [14]. The malware attacks, cyber-attacks, spams, spoofing, phishing, etc. in systems and databases are considered as malicious efforts and security threats that are claimed usually by the organizations. The network threats and security concerns may arise from all touchpoints of the users who collect and store data into databases such as sensors, RFID systems, telephone records, customers feedback on social networks, cameras, multimedia, and all other form of scientific data and calculations, that may be useful for the analysis purpose within an organization[3] [5] .

Additionally, unique characteristics are exhibited usually in Big Data systems having large volumes of data, which itself is a complex task to handle and control. In order to overcome this complexity, clouds are formed to manage data, although new challenges of security threats arise from cloud data management and processing [2] [6]. The cycle of data collection, integration, privacy, analytics, and security is threatened at every level. These challenges require a new network framework that is able to manage the acquisition of data, storage, transmission, and processing of volume of data as well as focusing on privacy and data security [4] [5].

A system architecture and framework is proposed for the stimulation of security systems in the Big Data and internet-based security. The factors show that in internet-based security, phishing has been one of the most common threats to security in big data systems. The techniques of data analytics in information security and Big Data are transforming over the years, therefore, new models and frameworks are proposed to secure the data from malwares. An effective proposed framework is Metamorphic Malware analysis and real-time detection or commonly known as MARD that helps in detecting real time information and Big Data security. It is useful because it helps in detecting malwares and threats, and enables the system for the preservation of data by the technique of anonymization of data from both bottom-up generalization and top-down specialization techniques[11] [13].

IV. PROPOSED METHODOLOGY

The methodology of this research is based on the semantic gap analysis of the existing security frameworks, to be tested on the platforms of Big Data systems through the proposed contemporary framework. The examination and investigation of data availability, confidentiality and integrity will be done.

In this proposed framework, a novel strategy is introduced to investigate the network traffic activities in the environment of Big Data systems. The proposed framework examines suspicious activities and malignant information travelling through the networks by utilizing Hive Queries integrated in an HDFS environment.

The objective of this proposed framework is to observe the extraction of data through a K-means algorithm and to operate the classification of the information by utilizing Support Vector Machine, Hive allows working in queries enabling to group and set an order of datasets. The details of this proposed framework and the complete process of network security analysis will be discussed further in the data analysis discussion. Following Figure 1 shows the network security analysis framework to be used in the Big Data environment.

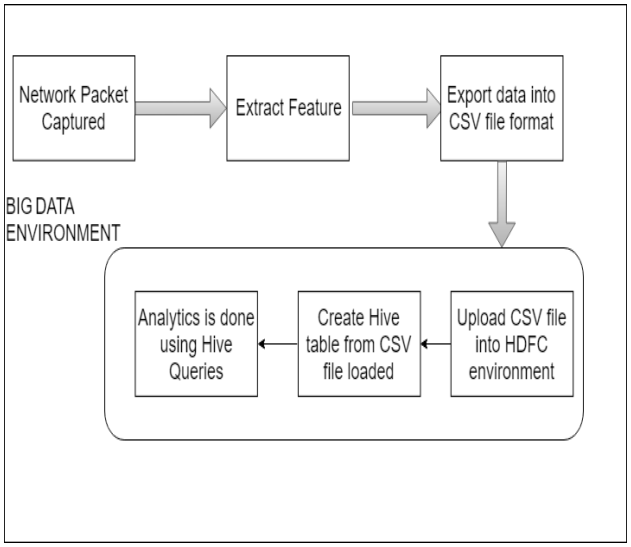

Figure 1: Network Security Analysis

It allows a more detailed analysis of the data, allowing simultaneous access to its properties as well. This framework procedure is then processed, extracting the metrics of the contents produced through operations. These metrics are then analyzed; detecting queries and identifying malicious operations existing in the network. Activity information will be loaded into the Hive database in Hadoop Distributed File System (HDFS) where it is examined and is sorted IP Wise, Port Wise and Protocol Wise, as shown in Figure 2. K-means algorithm is usually used on the information with missing labels and utilizes the concept of classifying groups based on the variable K. Data points are assigned to these groups based on their similarity in characteristics and features. Moreover, to find out the semantic gap, the semantic structure is studied as a new and novel approach and the dataset is used for the gap analysis.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

1. The issues and gaps in data analysis will be identified and envisioned using Hadoop technological system. The proposed architecture uses an iterative refinement method to process the result and extract feature wise metrics of information. The classification of datasets is proposed to be based on the algorithms of Support Vector Machine (SVM). The method processes the information for which labels have been assigned, distinguishes and identifies different groups, classifying them into different entities based on their assignment of data. This narrows down the data into specifics, which helps in the detailed assessment of the availed information, characterizing them depending on their operation. In the first phase, the data of network traffic is captured by the assistance of Wireshark. The dataset of NCCDC which is captured by utilizing the application of Wireshark is present in the format of PCAP. Significantly, Wireshark contains a programming interface application for capturing packets from a network. Different features are developed from the files of PCAP like window size, flag, protocol length, destination, sources ,timestamp and packet number as shown in the table, the data analysis focuses on the process of detecting Denial and Null, SYN scan, SYN/FIN, SYN Flood of various attacks of services by implementing the Hadoop technologies in Big Data.
2. In the second phase, the data of PCAP with a various extracted features use the customized file which is present in the application of Wireshark, This data further exported to .csv format for enabling the data exportation into the application setup of Hive database in the environment of HDFS.
3. In the third phase, the data present in .csv format is uploaded into environment of HDFS. Additionally, this data can be exported or be uploaded into environment of HDFS with the help of Hue services and Command line interface.
4. The fourth phase involves the process of uploading the data into the environment of HDFS, which is introduced in the table-log of data formatted and developed in the database of Hive .The table scheme of log data is presented in Table 1.

.

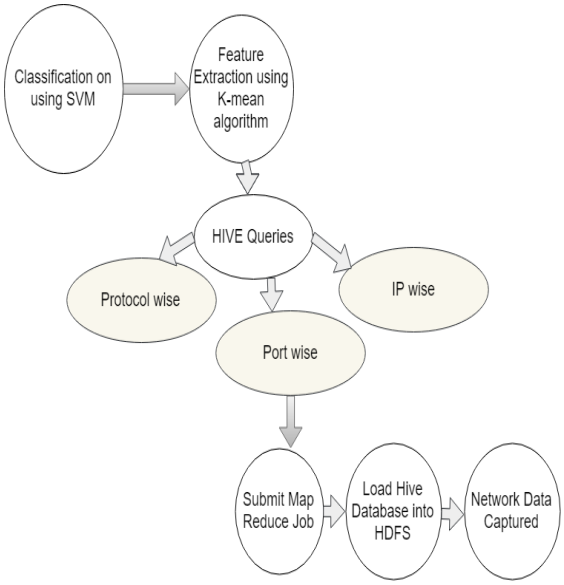

Figure 2: Proposed architecture

In the fifth phase, a study is performed by the assistance of Hive queries for identifying attacks on the network which is demonstrated and further explained in the section Table 1shows the parameters extracted from the network traffic and table 2 shows the log data type that is inserted into the hive, The .csv file consists of a flag field present in the hexadecimal format. This flag is converted into decimal format during the process of dumping data into the database of Hive. Furthermore, queries are developed for detecting Service Denial, Null scan, SYN scan, SYN flood attacks ,significantly the ones which are carried out on the protocol of TCP. In order to develop these queries for detecting the attacks, the process involve the between the attacks on theTCP protocol and its features

Table1: Parameter extracted from network traffic.

Parameter	Description
No	Number of the packet
Flag	Modified column
Source	IP address: packet transmitted
Destination	IP address: packet transmitted at end
Time	Timestamp of packet
Protocol	Protocol associated with the packet

Table 2: Log Data Type Into Hive Table

Column	Data type
Packet no	int
Flag	int
Source	string
Destination	string
Time	time

Protocol	int

Table 3 shows the correlation between various types of attacks and different parameters present in the TCP protocol.

Table 3: Types of attack on TCP Protocol based on Hive queries

Types of attack on TCP Protocol	Parameters
XMAS Scan	FIN,PUSH TCP Flag41
Null scan	TCP flag=0
SYN Flood	SYN &ACK,TCP flag=19

V. RESULTS AND OBSERVATION

This section presents the experimental setup and its result. Initially the data is taken from a cyber security dataset and sorted in HADOOP platform 2.7.0.2 with 64-bit platform and 8GB RAM .The data size (NCCDC dataset) is 97 MB and it contains approx. 9.7 million packets. The entire data packets are exported to Hadoop Distrusted File System and Hive queries are created. These queries import the packets and investigate them to identify the SYN Flood, Xmas Scan, and Null Scan attacks. The experiment is conducted to observe the datasets from a bundle stream and analyze them with the proposed idea.

A. *Hive Queries used for attack Detection*

Csv file format in hexadecimal format is converted into decimal format using Hive queries. Queries are designed to detect denial of Service, Null scan ,SYN Scan, mostly attacks which happen on the TCP protocol.

Due to a half -open connection or by leaving no port on the server, service request for the client is processed . A SYN packet is send to the server by a new client machine, and the server acknowledges its by sending a SYN-ACK(Acknowledgement) packet back to the client, when a SYN-ACK packet is sent to the client from the server, TCP flag is set to '18'If the client responds to the ACK packet, TCP flag is set to'16'.

Following is the Hive query used to detect a SYN Flood attack :

Select a.dest as Source_ ip,
a.source as dest_ip, count (dest_ip) from
tcp set flag = 18

B. *XMAS Scan*

If the ports on the server are closed, Xmas scan is revised by the attackers.

Hive Query :for the detection of X-MAS Attack:

Select source a, count(ip) as Packets
from log data where flag = 4.

C. *NULL SCAN*

The packets with no flag set are transmitted in NULL SCAN. When these packets are transmitted to the open ports of the targeted machine, they are rejected.

Hive query for the detection of Null Scan Attack:

Select source a, count(ip) as Packets from
Log data where flag = 0

In the sample dataset, a,SYN Flood detection query was executed and 1008 malicious packets were found. 86 and 761 malicious packets resulted ,when Null Scan detection queries and XMAS scan detection queries were performed respectively ,the percentage of malicious attacks is shown in the figure 3.

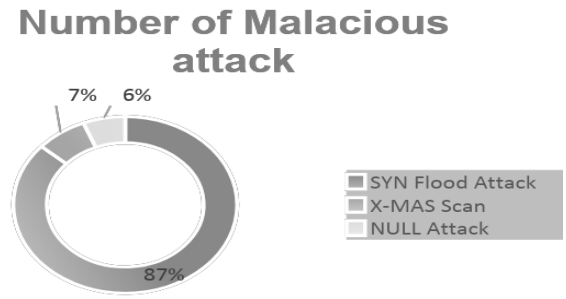

Figure 3: Number Of Malicious Attack

The results which are represented in Figure 3have shown that these queries have achieved 56% improvement in the analysis of the complexity in big data platform.It has been observed that the total time taken to identify threats and suspected contents drastically improves when we integrate K-algorithm and SVM algorithms, to determine and assign the datasets to Hive queries processes in an HDFS environment.

Following is the set of data achieved through the proposed framework procedure, classifying the information into Protocol, IP and Port Wise to compare, assess, inspect and deduct any anomalies existing in the network. The proposed framework has given positive results by identifying data packets that may be malicious and lead to network attacks and system intrusion. On the other hand, this proposed framework has few limitations as well. The limitations are related to the time taken for identifying and reporting of threats when the data packets are large and in huge volume. Moreover, the data analysis used Hadoop and Big Data technologies in order to run the queries on network packets and large volumes of data.

Figure 4 shows the reduction in dataset when the total number of packets are 715400 and marked packet are 377185.

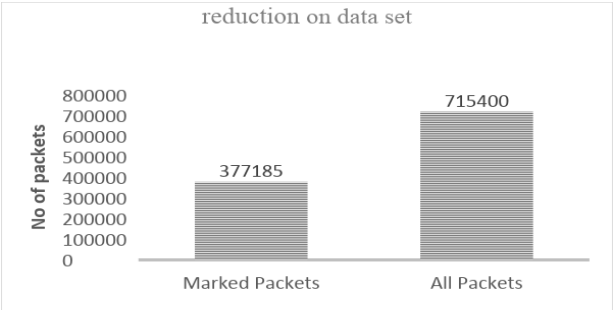

Figure 4: Reduction on data set

The time taken to identify threats needs to be reduced for more effective and positive outcomes. The response time should always be less in real-time attack situation because lower the response time, the identification of the connection of the attackers and overcoming it become easier.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The total time taken by the old technique that used only Hadoop and Hive queries in data analysis is more than the proposed technique. Figure 5 shows the comparison of the existing technique vs the proposed technique. The usage of the K-mean clustering algorithm helped in developing a relationship between network protocols and systems. The patterns of different types of attacks could be detected through this method. Furthermore, a two-tier architecture can also be focused for the supervision of threat detection. The anomalies can be grouped into clusters based on the similarities and differences between them.

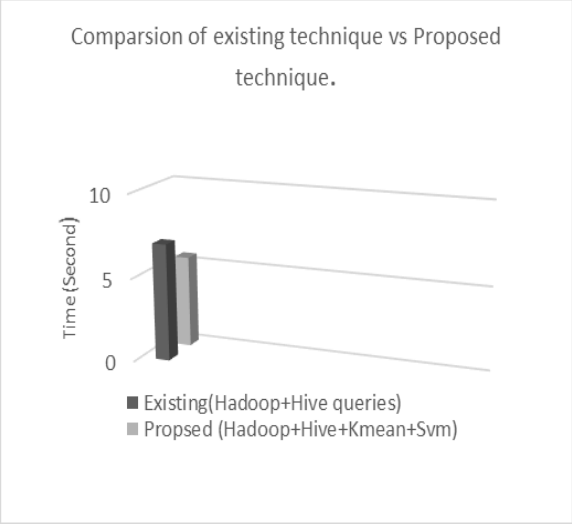

Figure 5 :Result Shows Proposed Framework Perform Better Than Existing.

VI. CONCLUSION

This paper proposes a modified framework for conventional as well modern security management. The purpose of this framework is to work in the HDFS environment, integrating with Hive queries and additionally utilizing the K-algorithm and SVM algorithms to improve the existing frameworks in order to efficiently implement the Big Data network security analysis.

The results have shown an improved efficiency in the data analysis process. Through this proposed idea the time taken to analyze large volumes of information is reduced significantly. There can be adaptations of various applications through this proposed framework in the future. The positive observations have encouraged this proposed framework to be considered for the enhancement of security analysis in Big Data network.

Moreover, the findings from the experiment shows some prevailing points that can help the professionals in demonstrating the implementation of a robust, sensitive, and a scalable system of Big Data. Furthermore, the findings and observations show that the existing platforms and frameworks are not as effective as the proposed one. The new framework will be helpful in the identification of anomalies and gaps in datasets, which will further lead to a successful mitigation of security threats and intrusions in the system.

REFERENCES

1. Naik, N., Jenkins, P., Savage, N. and Katos, V., 2016, December. Big data security analysis approach using computational intelligence techniques in R for desktop users. In Computational Intelligence (SSCI), 2016 IEEE Symposium Series on (pp. 1-8). IEEE.
2. Bachupally, Y.R., Yuan, X. and Roy, K., 2016, March. Network security analysis using Big Data technology. In SoutheastCon, 2016 (pp. 1-4). IEEE.
3. Lämmel, R., 2008. Google's MapReduce programming model—Revisited. Science of computer programming, 70(1), pp.1-30.
4. Big Data Basics-Part 5- "Introduction to Map Reduce" https://www.mssqltips.com/sqlservertip/3222/big-data-basics--part-5--introduction-to-mapreduce/
5. HADOOP Map Reduce Tutorial- https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm
6. Baijin Yang, Tonglin Zhang " A Scalable Meta-Model for Big Data Security Analyses" IEEE 2016
7. Yuangang Yao, Lei Zhang, Jin Yi, Yong Peng, Weihua Hu, Lei Shi " A Framework for Big Data Security Analysis and the Semantic Technology" IEEE 2016
8. Tiwari, H.C., and Yadav, S. 2015. A review on big data and its security. International Conference on Innovations in Information. Embedded and Communication Systems 1-5.
9. https://wiki.apache.org/hadoop/DataNode
10. Andrea Trevino, Introduction to K-means clustering (12.06.16) https://www.datascience.com/blog/k-means-clustering
11. Dean, J. and Ghemawat, S., 2010. MapReduce: a flexible data processing tool. Communications of the ACM, 53(1), pp.72-77.
12. Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P. and Murthy, R., 2009. Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment, 2(2), pp.1626-1629.
13. Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Zhang, N., Antony, S., Liu, H. and Murthy, R., 2010, March. Hive-a petabyte scale data warehouse using hadoop. In Data Engineering (ICDE), 2010 IEEE 26th International Conference on (pp. 996-1005). IEEE.
14. Lan, L. and Jun, L., 2013, December. Some special issues of network security monitoring on big data environments. In Dependable, Autonomic and Secure Computing (DASC), 2013 IEEE 11th International Conference on (pp. 10-15). IEEE.