Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 2, Number 10; April-June, 2015 pp. 869-873

http://www.krishisanskriti.org/jbaer.html

Botnet Detection Techniques with Data Mining using MapReduce

Garima Kaushik[1], Samruddha Patil[2] and Tanya Chawla[3]

[1]Graduate (B.Tech-CSE) BPIT, GGSIPU, Delhi
[2,3]Student (B.Tech-CSE) BPIT, GGSIPU, Delhi
E-mail: [1]garimakaushik13@gmail.com, [2]samruddha1401@gmail.com, [3]tanya.chawla94@gmail.com

Abstract—*The high insistence for management of big data has led to creation of Hadoop and Map Reduce. These are quintessential applications when incorporated with botnet detection techniques in order to attain the information of the corrupted nodes in a clustered architecture. Botnet is considered among the one of the most rampant and major threat for the legitimate and important computer networks and has been a fundamental issue for network security and data integrity which needs to be addressed at the highest priority. We survey various botnet detection techniques which are in association with the map reduce algorithms required to operate on big data. We further analyze data mining algorithms and observe their operation. Data mining algorithms have a fundamental characteristic that they allow detection over large amount of data which is mostly automated as compared to traditional approaches*

Most useful data mining techniques includes classification, correlation, clustering, aggregation for efficiently knowledge discovery about network flows and statistical analysis. Data mining techniques proven substantial and there are various attribute analysis which makes data mining algorithms convenient and effective for analysis.

Keywords: *BigData, Hadoop, HDFS: Hadoop Distributed File System, MapReduce, Botnet, Data Mining*

1. INTRODUCTION

The present big data scenario requires extensive and reliable hardware and software for proper processing and storage applications such as Hadoop, HDFS and MapReduce.

This gradual shift from enterprise to clustered architecture database storage has opened new possibilities for the operation of data mining techniques such as decision trees, clustering and classifications rules along with the functionality and benefits of MapReduce.

Since the clustered architecture is based on a number of nodes or computers collocated and associated with each other in a stack, it becomes essential to maintain the security and integrity of the data stored and employ the protocols required for detection of threat in the entire database.

Botnets have been a fundamental issue for network security in the recent years and has been affecting the performance of major government, business and commercial computer networks. A botnet is considered to be in effect when a passive adversary is responsible the attack on any essential client with the help of various compromised computers and nullifies the risk of being captured by backtracking.

This paper is divided into six sections. The section 2 includes Botnets followed by section 3 which has Big Data. Sections 4 and 5 consist of Overview of Hadoop and Botnet detection using Data mining and MapReduce. In the last section we have concluded the paper.

2. BOTNETS

An accumulation of interconnected programs based and associated with the internet performing a similar task is referred to as Botnets. This in itself is very generalized form of attack as it can occur as spamming emails , monitoring Internet Chat Relay(IRC) or perform denial of service to a legitimate user.

Botnets have their own organization and have independent structure. Botnets have characteristic to employ a Command and Control (C&C) strategy for its operation. A network which is compromised is comprised of several hosts or bots which receive instructions to perform a coordinated tasks from the botmaster or passive adversary present in the network itself as a legitimate server. The major medium to take control of legitimate users is through worms and Trojan Horses which are installed in the victims computer through spam of malicious softwares. The victims known as bots or zombies have no or little knowledge of them being a part of a botnet. Botnets can have either centralized architecture where one or more botmaster is responsible for the Command and Control (C&C) or distributed architecture where all the bots can be considered as the botmaster hiding its identity efficiently [1][2].

2.1 Threats of Botnet

Kind of botnet attacks are [3]:

i) Email Attack: Email social engineering attacks, promote user to open an attachment and follow an unsolicited link. When link or file exploits system becomes directly infected with malware. These attacks are commonly combined with phishing attacks that attempt to coerce the user into providing sensitive information.

ii) Web Client Attack: This is the technique to spread malware through Web. The victim is lured to malicious web sites, which are under the control of attacker and various techniques can be employed in order to cripple the security of browser of the victim. A malware is installed in the system if this process is a success and the user does not have its information.

iii) Instant messaging attacks: A passive adversary creates a malicious IM account and sends unreliable instant messages to genuine users. These messages appear to be authentic but leads the user to websites which initiate download of untrustworthy content or files

iv) Distributed denial of service attacks: Current bot variants include the ability to participate in distributed denial of service (DDOS) attacks against internet targets for revenge or profit. Distributed Denial of Service (DDOS) attacks are a type of Denial of Service (DOS) attack where several compromised infected computer systems are used to target a single computer system. This targeted computer system suffers from Denial of Service attack which means that legitimate users who are assessing the server or system cannot do so due to excessive number of requests on it. The motive behind a DDOS attack can vary from temporarily denying the service of a server to its legitimate users i.e. causing the server to slow or shut down to sending huge amount of random data to use up target computer system or server's bandwidth. There are several types of attacks that come under DDOS attacks such as TCP connection attacks, Volumetric attacks, Fragmentation attacks and application attacks. The intention of a hacker is to launch an attack and remain anonymous so that the source of the attack remains undetected. This poses a serious threat to any online server whose services are already being used at a significant level.

v) Click Fraud: Click fraud takes place when by illegitimate means visits are made to an online advertisement or other resource charged to the sponsor on a per-click basis. The clicks which are made by the compromised user are often results in sending web requests. The clicks lead to benefit to the advertiser and hence they are one of the primary medium for any botnet type attack to be performed.

vi) Key logging: Software key loggers capture keyboard events and record the keystroke data before it is sent to the intended application for processing. Data is capture by the spyware prior to encryption. Major targets are credit card information, authentication credential, email info.

2.2 Botnet Detection Techniques

There are various types of botnets targeting specific clients and networks depending upon their capability. In respect to this diversity following are various botnet detection techniques [4] [5]:

i) Active analysis: In this technique the information of the detection is provided to the bot master indirectly or directly. The bot malware is captured and malicious parts are deleted. Examples are Honeypots and Honeynets.
ii) Passive analysis: In this approach the analysis is done by monitoring the network traffic generated by botnets. Secondary effects of botnets are considered in this approach such as broken packets in the network data which is the result of DDOS attacks.
iii) Signature based Intrusion Detection System (IDS):
In this technique the information of the known signatures and intrusion patterns are compared with existing network traffic and initiates an alarm on discovery.
iv) Anomaly based analysis : In this technique the expected behavior of system is compared to the input data. This includes detection upon the number of anomalies in the network traffic.
v) DNS based analysis : In this technique the detection is done by observing the DNS traffic and DNS anomalies. However this technique is hard to implement on recent botnet detection.
vi) Data mining approach : The main goal is to detect useful patterns in an unlabeled data to determine regularities and irregularities. There are various types of Data mining detection techniques such as flow correction, classification, association rules, network flow, clustering.

Table 1: Comparison of Botnet Detection Techniques

Type	Unknown Bot Detection	Protocol & Structure Independent	Encrypted Bot Detection	Real-time Detection	Low False Positive
Signature based	No	No	No	No	Yes
Anomaly based	Yes	No	Yes	No	No
DNS-based	Yes	Yes	No	No	No
Mining based	Yes	Yes	Yes	No	Yes

3. BIG DATA

Till 2010, the term 'big data' was virtually unknown.2011 marked an epoch for big data as it was being widely touted as the latest technology. Today it has been adopted by everyone from outsourcing service providers to cloud services. The mammoth volume of structured and unstructured data that cannot be processed using conventional technologies and software is termed as 'Big Data'. Big data has come across as an eclectic collection of data points that can be used to derive business values. Analysts define big data using a 3V model [6][7] as follows :

1. Volume: In today's time, Facebook ingests burgeoning amount of data nearly 500 terabytes everyday. Analysis of such huge amount of information is remarkable and related to Big data.

2. Velocity: It refers to the speed at which data points are generated and changes. Sensors embedded into everyday objects producing large log data in real time, dynamic stock trading algorithms are some examples.

3. Variety: Big data being a heterogeneous collection of data consists of various formats and structures. It may contain geospatial, audio and video files, metadata and unstructured text files.

4th Vital V: Value : The vital V of Big data solutions provides a semantic analysis technique as value can be added to the raw information by analyzing sentiments contained in it.

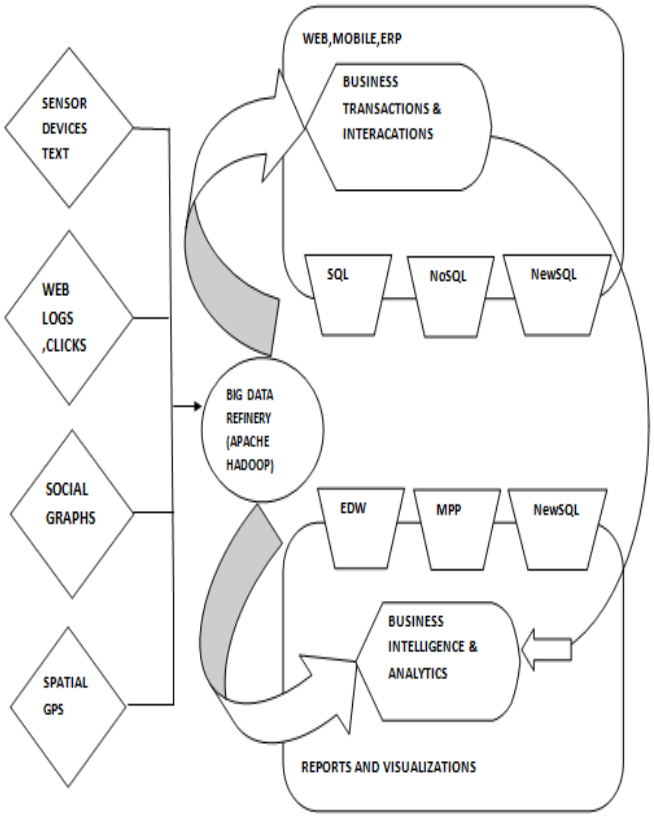

Fig. 1: Big Data Architecture

To sum up, it can be said that Big data benefice the organizations by providing them with a unique combination of existing data, transient data and externally available data to extract additional value

Big data is controlled mainly by 2 classes of technology: operational systems providing operational capabilities for real time work analysis, and analytical systems for providing capabilities for complex analysis [8].

Table 2: A comparative view of analytical and operational systems

Parameters	Operational system	Analytical system
Technology	NoSQL big data systems emerged to address a set of applications.	MPP and MapReduce have emerged as to overcome the shortcomings of traditional relational database .
Data Scope	Data scope is operational i.e. it provides an insight into trends derived from real time data with least amount of coding requirements.	Data scope is retrospective that provides real value to business.
Latency	1ms-100ms	1min-100min
Type of queries	Queries are quite selective when working with operational system.	Queries may or may not be selective in analytical system.
End Users	Customers can directly interact with operational system.	Data Scientists can work with the analytical system.

4. OVERVIEW OF HADOOP

In order to compensate with avalanche of the big data and the challenges related to it spreading over today's data storage scenario, Hadoop was created by Doug Cutting from Apache which allowed the systematic and compatible solutions to be implemented in various operations included in big data handling and management. An open source and parallel processing framework, Hadoop is basically dependent on clustered enterprise architecture instead of traditional enterprise architecture which works on Linux operating system. Various big data handling models have been developed subsequently over the years such as pig, hive and Hbase. After learning its capabilities Hadoop has been widely accepted and wanted in various multinational organizations which deal with big data and its applications such as Yahoo, IBM, Amazon and many more.

Hadoop's premier goal is to concealment of the intimate information for the parallel processing which comprises of computation done after the unification of results, distribution of data through parallel processing nodes present in the each cluster and to deal with failure of subtasks by restarting them without afflicting any risk on the integrity of data. Hadoop basically has two components Hadoop Distributed File System (HDFS) and MapReduce [9].

4.1 Hadoop Distributed File System (HDFS)

Based on the Google File System (GFS) presented by the Google Labs presented in 2003, the Hadoop distributed File System created by Yahoo is a big data storage model which allows the big data files to be stored in a clustered architecture befitted for parallel processing on a hardware which is inexpensive and available generally.

These clusters in HDFS contain a specific number of nodes which again comprises of three basic components processing cores, main memory which is attached to a set of commodity discs. The memory of a node is represented on disks but broken in the form of blocks of each 124 Mb or 64 Mb. Furthermore accompanied with abstraction, these blocks contribute storage of large files with simplicity and fault tolerance. In order to achieve high availability of data the HDFS maintains proper replication of data spread throughout the entire cluster.

The nodes follow a master-slave hierarchy and are classified into two basically Name-nodes (Master Nodes) and Data-nodes (Slave Nodes). In each cluster there is one Name-node which is responsible for managing the various Data-nodes in that respective cluster. Furthermore Name-node also manages the metadata and file systems namespace present in the data-nodes and gives out directions to provide various operations such as opening, closing and renaming new files and directories. Data is not stored nor passes through the Name-node. It also the Name node which is specified the largest ram possible in the entire Hadoop system. There is a Secondary node present which occupies the duties at the time of Name-node failure.

The responsibility of the DataNode is to manage the blocks and to deliver the data to the client. A DataNode stores blocks from different files and reports the list of blocks it has to the NameNode while requiring less cost and memory with no hardware replication. Another important functionality of Name-nodes is Job Tracker Node which is responsible for the management of the jobs specified by the clients. It also directs the task to the task tracker where the data actually stored. It also works in close coaction with the Task Tracker which manages the jobs required in the Map Reduce. The NameNode is indicated of the activation of a DataNode after every heartbeat [10].

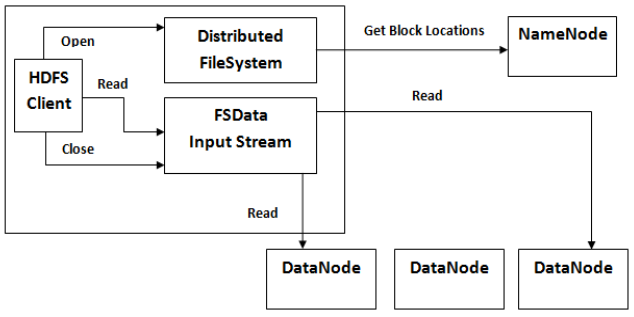

Fig. 2: HDFS Architecture

The Data flow in the HDFS basically dependent upon the nodes and their operations. The client which wishes to perform a read write or close operation first refers to the NameNode which provides the location of the DataNode where the data is stored in the block. If the data which is to be read or write upon is stored in multiple blocks among different data nodes the Distributed File System acquires the location of next DataNode which has the required block and accesses it. All these activities are transparent to the user and hence it reads or writes the data in a continuous flow. When the data is no longer found in the blocks at the DataNode, the operation is terminated [10] [11].

4.2 MapReduce

Inspired by Google's Map Reduce which is used to search and extract large data in web applications, Hadoop MapReduce has an architecture which supports and implements the management of large data sets structured in a clustered database [12]. It is essentially collection of programs which describe the duties such as re-executing a failure, monitoring and scheduling of any task .These programs can be written in various languages such as C++, Python, Ruby and Java. MapReduce provides the privilege to programmers to abstain from the complications accompanied by the parallelization of large data in a clustered database. The operation of MapReduce is classified in two segments Map and Reduce.

In the first segment Mapper function is applied to convert each data elements provided in a list into an output data elements in pairs of a key and a value represented as {key,value}. This list of data elements is also sorted and filtered so that to provide a data with proper structure and maintain its integrity. In the second segment the Reduce function is applied in order to accumulate all the input data elements to produce a singular output which is as per the requirement of the client. In this segment each Reduction task is responsible for the output of {key,value} pairs which is the result of processing of data from a single key taken as an input once at a time.

Hadoop allows MapReduce programs to run by allowing the clients to present a job to the JobTracker. The division of the job is the task of the JobTracker which maps and reduces it number of times until the MapReduce program is befitted to run on the query. It then designates the job to a number of TaskTrackers for further processing of the data and reporting the progress to the JobTracker which keeps a log of all jobs. The JobTracker remains active until the Job is finished [13].

MapReduce is also supports various scheduling algorithms such as FIFO Scheduler, Fair Scheduler, Capacity Scheduler, LATE-Scheduling and Delay Scheduling.

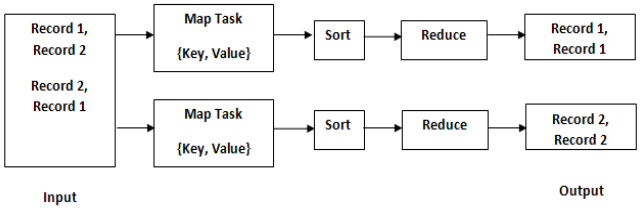

Fig. 3: MapReduce

5. BOTNET DETECTION USING DATA MINING AND MAPREDUCE

Data Mining is the process of finding useful patterns or correlations among fields in large datasets and for discovering regularities and irregularities in them. These techniques can be used for optimization purpose. The goal is to extract knowledge from an existing dataset and transform it into human understandable structure. Form a network log file essential data can be extracted using this. The most beneficial techniques for data mining clustering, aggregation, correlation and statistical discovery which can be utilized to determine knowledge from unlabeled data[14].

Clustering is a technique under data mining in which there are data points having feature values and a common measure are clustered together. The algorithms under clustering find structure in unlabeled data and is considered as unsupervised learning. We further go into details of two clustering algorithms, K-means and expectation maximization (EM) [15]. K-means algorithm is a clustering algorithm which partitions all data instances into k clusters maximizing intra-cluster similarity and minimizing inter-cluster similarity. New cluster centers are calculated by using mean for each cluster and then reassigning all instances to the cluster whose center is closest to it. It is repeated till there is no inter-cluster movement. It is a computationally difficult (NP-hard) problem. These are quite similar to the expectation maximization algorithm and both use cluster centers to model the data.

One of the biggest challenges includes adding structure to the unstructured data, which is done through natural language processing.

As the demand for parallel computing is increasing, cloud computing techniques can be used with data mining to achieve greater efficiency. MapReduce provides good parallelism and with it, data mining can enter a new era of implementation [16].

The traditional investigative methods based on signature are not sufficient as compared to automated detecting features of data mining when large amount of data is operated upon and analyzed [17].

6. CONCLUSION

An essential part for management of any data is to provide security to the highest degree and measures for safeguarding the integrity of the data. In recent times since the demand for storage of BigData in clustered and parallel architecture is on the rise. This has led to consider the subject of its security to be considered as topmost priority.. The introduction of new technologies to process BigData such as Hadoop and its features such as HDFS and MapReduce are prone to already existing threats present in the database and the network scenario such as botnets. However the flexibility and compatibility of these latest softwares has allowed various features of relational data warehousing such as data mining and its techniques to merge with MapReduce and result in the formation of botnet detection techniques. These detection techniques can be relied on to counter various botnet attacks and refrain from compromising the data stored in the clustered architecture. Furthermore this opens another area of research and analysis of the various technologies which can be collaborated with Hadoop and enhance its security and protect the data from malicious sources and cyber-attacks.

REFERENCES

[1] Jing Liu, Yang Xiao,Kaveh Ghaboosi, Hongmei Deng and Jingyuan Zhang, " Botnet : Classification, Attacks, Detection, Tracing and Preventive Measures", EURASIP Journal on Wireless Communications and Networking, Volume 2009, Article ID 692654.

[2] Jody R. Westby, "Legal Guide to Botnet Research", Paperback, 2014

[3] Paul Barford, Vinod Yegneswaran, "A look inside botnets", Advances in Information Security Volume 27, 2007, pp 171-191

[4] Erdem Alparslan, Adem Karahoca and Dilek Karahoca, "BotNet Detection: Enhancing Analysis by Using Data Mining Techniques", September 12, 2012.

[5] Alireza Shahrestani, Maryam Feily, Rodina Ahmad, Sureswaran Ramadas, "Architecture for applying data mining and Visualization on network flow for botnet traffic Detection" , International Conference on Computer Technology and Development,IEEE, Pages 33-37 2009.

[6] David Lauzon, "Introduction to BigData" , 2012

[7] Sam Madden, "From Databases to Big Data" , IEEE Internet Computing 16(3): 4-6, May-June 2012

[8] Bart Baesens , "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications", Wiley,2014.

[9] Kai Ren, Yongchul Kwon, Magdalena Balazinska, Bill, "Hadoop's Adolescence: An Analysis of Hadoop Usage in Scientific Workloads", Howe. 39th International Conference on Very Large Data Bases 2013 (VLDB 2013), August 26-30, 2013. Proceedings of the VLDB Endowment, Vol. 6, No. 12, 2013

[10] Rakesh Varma, " Survey on MapReduce and Scheduling Algorithms in Hadoop", International Journal of Science and Research, Vol.4, Issue 2, February 2015.

[11] Tom White, "Hadoop: A definitive guide", O'Reilly, Farnham", October 2010.

[12] Dean, J. and Ghemawat, S., "MapReduce: a flexible data processing tool", ACM 2010.

[13] Jaliya Ekanayake, Shrideep Pallickara, and Geoffrey Fox, "MapReduce for Data Intensive Scientific Analyses", Fourth IEEE International Conference on eScience,2008.

[14] Alireza Shahrestani, Maryam Feily, Rodina Ahmad, Sureswaran Ramadass. "Architecture for applying data mining and visualization on network flow for botnet traffic detection", International Conference on Computer Technology and Development,IEEE, Pages 33-37 2009.

[15] Huan Lai, "Applying Data Mining Techniques to MapReduce", Techblog May 2010.

[16] Viki Patil, V. B. Nikam "Study of Data Mining algorithm in cloud computing using MapReduce Framework", Journal of Engineering, Computers & Applied Sciences (JEC&AS) Volume 2, No.7, July 2013.

[17] Jignesh Vania, Arvind Meniya, H. B. Jethva, "A Review on Botnet and Detection Technique" International Journal of Computer Trends and Technology- volume4Issue1- 2013.