ANALYSIS OF INTRUSION DETECTION IN BIG DATA

Anmol Handa\textsuperscript{1}, Siddhant Gupta\textsuperscript{2} and Deepika Bhagwani\textsuperscript{3}
\textsuperscript{1,2,3} B.Tech (CSE), MAIT\textsuperscript{1,2,3}, New Delhi, India

Abstract – Big Data refers to technologies and initiatives that involve data that is too diverse, fast-changing or massive for conventional technologies. This paper describes the Big Data usage and the security issues associated with it. With Big Data databases, enterprises can save money, grow revenue, and achieve many other business objectives by building new applications, improving the effectiveness and lowering the cost of existing applications, realizing new sources of competitive advantage. The continuous collection of traffic data by the network leads to Big Data problems that are caused by the volume, variety and velocity properties of Big Data. This paper also focuses on the problem and probability of intrusion in Big Data. Intrusion refers to the hacking of the system or a network which poses a security risk to the private information of the organization. To address the problem of network intrusion, a network intrusion detection system is used which detects such intrusions through the analysis of traffic on the network to monitor signs of malicious activity. It also keeps an eye on the packets on a network wire and its main objective is to look out for discrepancies. This paper describes the various techniques for intrusion detection and tries to identify the best technique.

Keywords – big data, data security, intrusion detection, Hadoop, Machine Learning.

I. INTRODUCTION

Big Data is a large collection of data sets, so large in variety, volume and velocity that it is impossible to evaluate and handle them using on hand management tools or traditional methods. To obtain value from this data, you must select an alternative way to process this data. Big Data has become practicable as cost effective approaches have emerged to tame the velocity, variety and volume of massive data.

Volume: It is the feature of Big Data to process large amounts of information. It requires scalable storage and a distributed approach to access complex data.

Velocity: Velocity in Big Data refers to accelerating rate at which the data flows into an organization.

Variety: Data rarely presents itself in a perfectly ordered form which is ready for processing. A common pattern in Big Data systems is that the source data is too diversified and is not present in proper relational structures.\textsuperscript{[1]}

Big Data is big and messy. It is not only too big to process but also too big to transport anywhere. Even if the data is not too big to move, its locality can still be a major issue. It is consistently reported that 80 percent of the effort involved in dealing with Big Data is cleaning it up and making it more presentable.

As the complexity increases, the current techniques and technologies that are dealing with Big Data stop functioning. Due to these above mentioned reasons, few applications that suffer from Big Data problem are network traffic analysis, geo-spatial classification and business forecasting and network intrusion detection. These applications require efficient Big Data techniques to tackle the problem on fly so as to attain an effective and efficient Big Data system. The techniques like Cloud technology, HIVE database and Hadoop Distributed File System can together address the problems of Big Data. In this paper, few shortcomings and challenges related to Big Data classification and analysis as well as probability of the intrusion in Big Data is discussed. Moreover, major techniques to deal with intrusion
detection and discussed.

II. WEAK POINTS OF BIG DATA

Recent researches have proved that Big Data has limitations in many areas like meteorology, genomics, environmental research and physical simulations. Other related areas are Internet search, finance and business. Big Data, due to its high complexity, cardinality and continuity suffers from the following problems at different stages namely date capturing, storage, transfer, analysis, searching and sharing of data [2]

A. Scalability: We want to scale Big Data very rapidly and elastically wherever and whenever we want, even on the cloud if needed but at the same time NoSQL solutions like mongoDB or HBase have their own scaling limitations. [11]

B. Continuous Availability: Every organization in this era relies on Big Data to feed their essential revenue generating business applications. So, even high availability of data is not enough. The data must never go down. However, downtime is embedded in the RDBMS and other NoSQL systems. So, many industries suffer from lack of continuous availability of data and thus, face massive losses in their business.

C. Workload diversity: Big data is available in different formats, shapes and sizes and therefore rigid schemas cannot efficiently handle them. Instead, we need a need a more flexible design. The technologies should fit to our data and not the other way round.

D. Performance: In this competitive world where a microsecond delay can cost you a big loss, Big Data must be transported at extremely high velocities, no matter how much workload it has and how much scaled it is but at the same time data handling range of RDBMS and NoSQL solutions put a serious haul on performance.

E. Manageability: Big Data should be efficiently captured and stored in distributed file system to implement intrusion prediction and proper storage. Continuity and complexity add difficulties to the task of management. [3]

Therefore, network topology for a Big Data system must be designed in such a way that Big Data analysis can be handled in a cost effective way adding high performance, continuous availability, scalability and manageability to the system and at the same time, handling the workload efficiently. Thus, the proposed network topology includes the technologies like Hadoop Distributed File System, Public Cloud and Network Traffic Recording System. [4]

Fig. 1. Big Data Network Topology [1]
These modern technologies can deal with the cardinality problem in Big Data system. These can be integrated to build a robust and flexible network topology with abundance storage which can change dynamically based on Big Data processing requirements.

However, this model brings several challenges that have to be dealt with an efficient and effective manner. Major challenges that we face with this topology fig.1 are communication and security challenges.

A. Communication challenge: In computer networking applications, communication should be effective and its cost should be minimal. The challenge here is to optimize the communication cost while making sure that the data requirements as well as the storage requirements on cloud are satisfied.

B. Data Security: Security in Big Data is a major issue and thus tampering if data can be done easily. A major security breach in the system is that a hacker can attack the data easily when it is being streamed from Hadoop Distributed File System (HDFS) to Cloud Computing Storage System (CCSS). [4] Thus, effective machine learning techniques need to be developed so that the challenges faced by the existing topology are taken into account.

III. INTRUSION IN BIG DATA

When a malicious attack is launched against a system, the attack typically leaves evidence of an intrusion into the system. Each intrusion leaves a kind of track behind (for example, unauthorized software execution, misuse of administrative privileges, failed logins, file and directory access). Intrusion refers to the hacking of the system or a network which poses a security risk to the private information of the organization.

Big Data is all the buzz in today’s time but its security is still on the primitive level. It allows tons of suspicious data to be analyzed and used to detect intrusion. In an enterprise, it analyzes volumes of network traffic to unveil inside trouble and advanced persistent threats.

When this network architecture to support Big Data was created, the security was not kept in mind which is now creating problems for system defenders. A mundane Big Data deployment uses Hadoop Distributed File System to manage the cluster of computers used to collect massive amounts of data fed into it that can create major security challenges, especially, if they are using traditional security tools. Current data security vendors believe that Hadoop and distributed cluster security can be addressed with established security solutions such as firewalls and intrusion detection/prevention technologies but no matter how advanced these approaches are they are unable to fully secure Hadoop clusters and distributed file systems. The traditional security products were basically designed to protect a single database; therefore, they can’t be efficiently implemented in a distributed file system. When you put them on a bigger scale, they can possibly become a single spot of failure for the entire cluster.

In the proposed topology, security is weak and there is high probability of intrusion in data. A stranger at a remote location can easily attack the system when the data is being streamed between the Cloud Computing Storage System (CCSS) Server, Hadoop Distributed File System (HDFS) and Network Traffic Recording System (NTRS). The attacker can easily shut down the server and access the private information of the enterprise.

To address the problem of network intrusion, a network Intrusion Detection System (IDS) is used which detects such intrusions through the analysis of traffic on the network to monitor signs of malicious activity. It also keeps an eye on the packets on a network wire and its main objective is to look out for discrepancies. Intrusion detection and prevention systems are focused on identifying possible incidences of intrusion, logging information about them and reporting attempts. Analysis and Intrusion Detection in Big Data.
IV. MAJOR NETWORK INTRUSION DETECTION TECHNIQUES

IDS come in a variety of “flavors” and approach the goal of detecting doubtful traffic in different ways.

ID systems are needed because of the increasing number of attacks on major sites and networks, including those of the NATO, and the U.S. Defense Department and many others [5]. There are network based IDS (NIDS) and host based IDS (HIDS). In network based IDS, the individual packets flowing through a network are analyzed on the entire subnet, works in a indiscriminate mode and matches the traffic that is passed on the subnet to the library which contains the information of all the known attacks. Once the attack is sensed or anomalous behavior is identified, the alert can be sent to the super user or administrator. In host based IDS, intrusion detection is done on individual devices or hosts on the network. HIDS monitors the in-bound and out-bound packets from the device only and will notify the user or administrator if suspicious activity is detected. All IDS use one of the two detection techniques:

A. Statistical anomaly-based IDS

An IDS which is anomaly based will observe network traffic and compare it against established rules or baseline. The rules will identify what is “normal” for that network- what sort of bandwidth(or frequency) is generally used, what rules are used, what devices and ports generally connect to each other- and alert the administrator or any user when traffic is detected which is significantly different, than the established baseline. The issue is that it may lift a False Positive alarm for a rightful use of bandwidth if the baselines are not intelligently configured [6], so baselines should be configured quite intelligently with proper machine learning or representation techniques so as to reach high percentage of correct results.

B. Signature-based IDS

A signature based IDS will check packets on the network and compare them against a list of attributes or signatures from known malicious threats. This is analogous to the way most antivirus software discovers malware. Here, the issue is that there will be a delay between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that delay period your IDS would not be able to detect the new threat.

V. ANOMALY DETECTION TECHNIQUES

Anomaly detection is based on a host or network. Many distinct techniques are used based on type of processing related to behavioral model. They are [7]:

A. Statistical based:

Statistical-based systems (SBIDs) take an altogether different approach to detect intrusion. The
concept of the SBID system is straightforward: it determines "normal" network activity and then the traffic lying outside the scope of normal is taken as anomalous. SBID systems tend to learn network traffic patterns on a particular network. Traffic analysis process continues till the SBID system is active, so, assuming network traffic patterns remain consistent, the longer the system is present on the network, the more correct results it gives. By analyzing network traffic and processing the information with tortuous statistical algorithms, SBID systems look for deviation in the established normal network traffic patterns. All packets are given an anomaly score (indicating the level of abnormality) and if the anomaly score is higher than a certain minimum level, an alert is generated. The key to any Statistical based system is its talent to learn things and differentiate normal from abnormal network activity.

For better understanding- Suppose you have a dog (Spot) and you feed him daily when you come home. Daily, you fill the empty bowl you find when you reach home. One day you reach home to see that the food bowl is half empty. Depending on various aspects, this event could be very enigmatic and demands further investigation, or maybe something you don’t need to think about. Now suppose if you come home one day and the food bowl is full. In this extreme case, it would be reasonable to assume the extreme. Dog may be sick, Dog may have run away, or maybe he’s not hungry at all. Due to the level of aberration of the situation, further investigation is surely needed. This is how SBID systems are intended to work. The system learns what is usual by observing an activity over time. When an anomalous event occurs, the SBID monitors it and consequently generates an alert so as to prevent further damage to the system.[12]

It is of following types:

1. Proximity-based

Here, no preceding assumptions about the data distribution model are taken. It has exponential computational intricacy due to the need to compute the distances of all data records. Types of proximity-based algorithms

1.1. K-Nearest Neighbor (k-NN) algorithms:

The learning phase groups the training vector space to a priori known classes, calculates the k nearest neighbors of an unknown sample using a distance metric such as Euclidean distance. Then it assigns the unknown sample to the most familiar class among the k nearest neighbors.

1.2 K-means clustering algorithm:

It compare new sample to the k prototype vectors and choose the closest, it’s computationally more efficient compared to k-NN.

1.3 K-medoids algorithm

Each cluster center is represented by a centrally located (the medoid) point rather than a prototype point.

2. Parametric

Here data points are modeled using pre-selected random distribution model. This technique fits boundaries around specific percentage of the data irrespective of the sparseness of the outlying region. If the user has an analytical knowledge that the data set fits such a model, then precise results are obtained which relies on good spread for the data. Otherwise bound of normality may omit many normal points.

3. Non-parametric

There are no assumptions about the underlying data distribution in this case. It is more suited for smaller sample size. The number and nature of parameters is flexible. Its confidence bounds can be much wider than that for the parametric counterparts. Predictions are not possible outside the range of observations.
4. Semi-parametric
It combines the velocity and complexity growth advantage of parametric methods with the model flexibility of non-parametric methods.

B. Neural Network Approach
An artificial Neural Network consists of a collection of treatments to convert a set of inputs to a set of searched outputs, through a set of simple processing units, or nodes and connections between them. Subsets of the units are input nodes, output nodes, and nodes between input and output form hidden layers; the connection between two units has some weight, used to determine how much one unit will affect the other. Two types of architecture of Neural Networks can be distinguished:

Supervised training algorithms, where in the learning phase, the network learns the desired output for a given input or pattern. The well-known architecture of supervised neural network is the Multi-Level Perceptron (MLP); the MLP is employed for Pattern Recognition problems. [13]

Unsupervised training algorithms, where in the learning phase, and the network learn without citing desired output. Self-Organizing Maps (SOM) are popular unsupervised training algorithms. SOM are used for classification problems.

A good introduction to Neural Networks is available in books. The most important property of a Neural Network is to automatically learn / retrain coefficients in the Neural Network according to data inputs and data outputs. Applying the Neural Network (NN) approach to Intrusion Detection, we first have to expose NN to normal data and to attacks to automatically adjust coefficients of the NN during the training or learning phase. Performance tests are then conducted with real network traffic and attacks to keep a check on its performance and also to measure its accuracy. [8]

C. Machine learning approach
Anomaly detection detects unknown attacks based on inspection, but it sometimes raises high false-alarm and is limited by training data IDS. It corrected by using some machine learning techniques which protect host and network from intruders before the attack has been launched. The final model produced must have higher accuracy levels, which shows both a high detection and an extremely low false positive rate. Some commonly used machine learning techniques are described below. [14]

1. Pattern classification
Pattern recognition is the action to take raw data and divide them on the basis of category. The supervised and unsupervised learning methods can be used to solve different pattern recognition problems. In supervised learning, it is based on using the training data to create a function, in which each of the training data contains a pair of the input vector and output. The training task is to compute the approximate distance between the input–output examples to create a classifier (model). Once the model is created, it can classify unknown examples into a learned class labels. [14, 15]

2. Single classifiers
The intrusion detection problem can be approached by using one single machine learning algorithm. Machine learning techniques (e.g. k-nearest neighbor, artificial neural network, decision trees, support vector machines, self-organizing maps, etc.) have been used to solve these problems.[14]

3. Hybrid classifiers
In the development of IDS, the ultimate goal is to achieve the best possible accuracy for the task at hand. This objective leads to the design of hybrid approaches for the problem of intrusion to be solved. The main idea behind a hybrid classifier is to combine several machine learning techniques so that the system performance can be improved significantly. More clearly, a hybrid approach typically comprises of two functional components. The first one takes raw data (input) and generates intermediate results. The second one will then take these intermediate results as the input and generate the final
results. Particularly, hybrid classifiers are usually based on joining different classifiers, such as neuro-fuzzy techniques. [15]

Machine learning approach covers all the major techniques in one or different way either by using single classifiers or hybrid classifiers and thus is an efficient technology which can be implemented in intrusion detection by making use of best Analysis and Intrusion Detection in Big Data techniques available and combining them to give higher accuracy to the IDS.[14]

VI. SIGNATURE-BASED IDS

Signature detection searches network traffic for a series of malicious bytes or packet sequences. The major advantage of this technique is that signatures can be easily evolved and understood if we know what network behavior we are trying to distinguish. It may look out for particular strings to detect intrusion. The events generated by signature based IDS can correspond the cause of the alert. As pattern matching can be done easily and more proficiently, so the amount of power needed is minimal to perform the matching for rule set. For example if the system that is to be protected only communicate via DNS, ICMP and SMTP, all other signatures can be ignored for better security reasons. [9]

Limitations of signature based IDS are that they only detect attacks whose signatures are already kept in database; a signature must be created for every attack; and new attacks cannot be detected. This technique can be easily deceived because they are only based on regular expressions and string matching. Also signatures work well against only the fixed behavioral pattern, they fail to deal with attacks created by a worm or human with self-modifiable behavioral characteristics. [9]

VII. BEST SUGGESTED TECHNIQUE

As we can see from above techniques discussed about the intrusion detection in big data that there are two major type of intrusion detection systems namely Anomaly based IDS and Signature based IDS. Out of these anomaly Detection is a better IDS as compared to Signature based IDS as it’s always ready to detect new type of malwares and intrusions as compared to signature which only detects the intrusion which have their imprints (signatures) already present in the system. Also the IDS are of two type’s viz. Network IDS and Host IDS. Both have their own importance in possible attack detection and prevention and thus a best intrusion detection System should be implemented at Network as well as at Host to develop a robust and secure system. Network based IDS tracks the individual packets flowing through a network and analyzes them on the entire subnet whereas Host based IDS monitors the in-bound and out-bound packets from the device only and alerts the user or administrator in case of suspicious activity. Therefore the best approach that can be implemented for intrusion detection in big data analytics is the machine learning approach which is capable of overcoming the basic shortcomings of anomaly detection techniques by protecting the network from intrusion before the attack has been launched. Therefore machine learning intrusion detection produces higher accuracy rate in intrusion detection and prediction and extremely low false positive rates. It uses pattern classification, single classifiers as well as hybrid classifiers techniques to detection anomalies in the network as well as host. This machine learning technique with the help of these classifiers covers all the major intrusion detection techniques namely neural network, statistical based with higher accuracy results and thus is considered best in the lot. Machine learning technique uses decision trees and genetic algorithms to develop programs as well as algorithms for intrusion detection. The genetic algorithm software package was a logical platform from which we could handle the difficult problem of intrusion detection. The use of decision trees for rule generation was made to serve as a deterministic alternative to genetic algorithms. [10] Machine learning methods are immensely superior in analyzing prospective customer churn across data from multiple sources such as CRM transactional and social media sources. High performance machine learning can analyze all of a Big Data and not only its sample. This scalability not only allow predictive solutions based on sophisticated algorithms to be more correct, it also shows the
importance of software’s speed to interpret the billions or trillions of rows and columns in real-time and to analyze live streaming data. [11]

VIII. CONCLUSION

We first gave a brief introduction to big data about what it is, where it is deployed and what are its uses. So, Big Data is a large collection of data sets, so large in variety, volume and velocity that it is impossible to evaluate and handle them using on hand management tools or traditional methods. Big Data is big and messy. It is not only too big to process but also too big to transport anywhere. Then we discussed the major issues related to big data and its various shortcomings related to Scalability, Continuous Availability, Workload diversity, Performance, Manageability. Then we discussed about the network topology best suited for big data which can deal with all these weak points effectively. After that we gave a brief about intrusion—what it is, how it happens, how it can be detected accordingly. We gave a number of techniques for intrusion detection and chose the best among the lot. Machine learning technique due to its wide variety of features and number of algorithms is considered best for intrusion detection in big data analytics.

REFERENCES