# SURVEY ON INTRUSION DETECTION ON NETWORK USING BIG DATA ANALYTICS WITH HADOOP

*Y. S. Kalai Vani[1]*

*[1]Assistant Professor, Department of Computer Science, Sindhi College of Commerce.*

**ABSTRACT**

**BACKGROUND**

Big Data is an emerging paradigm applied to datasets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. Big Data describes data sets that are too large to unstructured or too fast changing for analysis. Big Data analytics is the process of analysing and mining Big Data. Due to increase in number of sophisticated targeted threats and rapid growth in data, the analysis of data becomes too difficult. In this review paper, we are discussing about technique how Big Data is analysed by using the technique of Hadoop and why the Big Data Security Analytics is important to mitigate the security threats to secure the enterprise data more efficiently.

**KEYWORDS**

Big Data Analytics, Hadoop, Map Reduce, Big Data, Security Analytics, Targeted Attacks.

**HOW TO CITE THIS ARTICLE:** Vani YSK. Survey on intrusion detection on network using big data analytics with Hadoop. J. Technological Advances and Scientific Res. 2017;3(1):14-16, DOI: 10.14260/jtasr/2017/05

## BACKGROUND

Most attacks on the internet consist of opportunistic attacks rather than attacks targeted for some specific entity. An opportunistic attack is when an attacker targets various different parties by using one or various generic ways to attack such parties in the hope that some of them will be vulnerable to attack. In an opportunistic attack, an attacker will have a large number of targets and will not care that much on who the victim is, but rather, on how many victims there are. A targeted attack is much more effective and damaging for the victim since the actions performed by the malicious hacker are tailored. This means that it is much more difficult to stop a targeted attack than an opportunistic one simply because the attacks themselves are not general. With the huge amount of processed data available on internet, hackers also become so active with malicious attacks. Hackers target the analysed data and create(1) threats to their target environments for information.[2][3] Big Data security analytics is used for the growing practice of organisation to gather and analyse security data to detect vulnerabilities and intrusions.[4] The aim is here to make use of Big Data techniques to analyse the data and apply same to implement enhanced data security mechanisms. To obtain data for such systems, organizations pick a variety of hosts with a range of Security Analytics Sources (SAS). It is a system that generates messages or alerts and transmits them to trusted server for analysis and action. It can be Host-based Intrusion Detection System (HIDS), an antivirus engine that writes a syslog or interface that reports events to remote service, e.g. Security and Information Event Monitoring (SIEM) system. The malicious[5] and targeted attacks have become main subject for government, organisation or industry.[6]

*Financial or Other, Competing Interest: None.*
*Submission 28-12-2016, Peer Review 11-01-2017,*
*Acceptance 03-02-2017, Published 28-02-2017.*
*Corresponding Author:*
*Dr. Y. S. Kalai Vani,*
*No. 5, "Akshaya Residency",*
*31/4, First Main Road,*
*Shiva Shankar,*
*Hebbal, Bangalore-24*
*E-mail: kalaiys@rediffmail.com*
*DOI: 10.14260/jtasr/2017/05*

A subset of threats is Advanced Persistent Threats (APT), which are well resourced and trained adversaries that conduct multi-year intrusion campaigns targeting highly sensitive economic, proprietary or national security information. Their aim to keep their persistency without getting detected inside.

### Big Data Analytics and Detection Techniques

Big Data can be used to build more practical and successful Security Incident and Event Management Systems (SIEM), Intrusion Detection System (IDS) and Intrusion Prevention System (IPS).

### Detection Techniques

There are two basic categories of intrusion detection techniques- anomaly detection and misuse detection.

**A. Anomaly Detection**

Anomaly detection uses models of the intended behaviour of users and applications, interpreting deviations from this "normal" behaviour as a problem. 2-4 a basic assumption of anomaly detection is that attacks[7] differ from normal behaviour. For example, we can model certain users' daily activity (Type and amount) quite precisely. Suppose, a particular user typically logs in around 10 a.m., reads mail, performs database transactions, takes a break between noon and 1 p.m., has very few file access errors and so on. If the system notices that this same user logs in at 3 a.m., starts using compilers and debugging tools and has numerous file access errors, it will flag this activity as suspicious. The main advantage of anomaly detection systems is that they can detect previously unknown attacks. By defining what's normal, they can identify any violation, whether it is part of the threat model or not. In actual systems, however, the advantage of detecting previously unknown attacks is paid for in terms of high false-positive rates. Anomaly detection systems are also difficult to train in highly dynamic environments.

**B. Misuse Detection**

Misuse detection systems essentially define what's wrong. They contain attack descriptions (or "signatures") and match them against the audit data stream looking for evidence of known attacks.

5-7 One such attack, for example, would occur if someone created a symbolic link to a Unix system's password file and executed a[8] privileged application that accesses the symbolic link. In this example, the attack exploits the lack of file access checks.

There are different set of attacks in the cyber to overcome the problem in the cyber-attacks, many detection techniques are used. Many algorithms are based on the anomaly and misuse detection basis. Big Data analytics are latest technology, which is used to detect the cyber-attacks in an efficient manner and it hold huge volume of data in the fastest manner. In the below figure (1) explain the use cases of Big Data analytics in the efficient manner.

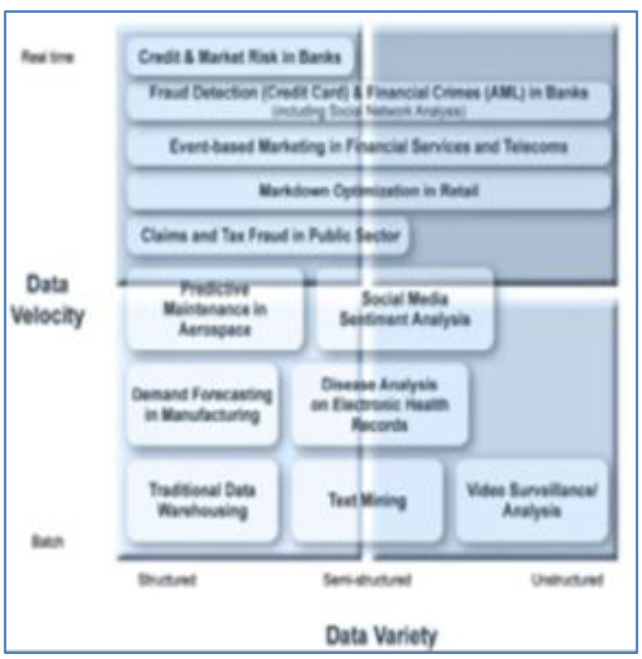


***Figure 1. Potential Use Cases for Big Data Analytics***

Big Data analytics is helpful in managing more or diverse data. It also helps to generalise new questions from observation, formulating new hypotheses, explore and discovery of new processed concepts and making decisions from testing. The main efforts done by Big Data analytic is the use of new analytics techniques on either new data or data that has been mixed in new ways. Big Data analytics used the tool Hadoop for processing the unstructured data. The main point is whether Hadoop will become as indispensable as database management systems. Hadoop has proven its advantages of use and cost where volume and variety are extreme.[10] Cloudera, Hortonworks and Map R are doing work for Hadoop on high-scale storage and MapReduce processing data into the world of analytics. Data analysis is a do-or-die requirement for today's businesses. Hadoop upstarts to traditional database players by analysis done by vendors.

**Hadoop- Tool for Analysis**

Hadoop is designed to process large volumes of information by connecting many commodity computers together to work in parallel in efficient manner.[7] The 1000-CPU (or processor) machines would cost a very large amount of money, far better than 1,000 single-CPU or 250 quad-core machines. Hadoop have tied these smaller and more reasonably priced machines together into a single cost-effective computer cluster. In a Hadoop cluster, data is distributed to all the nodes of the cluster present on which data can be loaded as shown in Figure 2. The Hadoop Distributed File System (HDFS) will do this distribution of large data files into chunks, which are managed by different nodes in the cluster.[8] An active monitoring system then re-replicates the data in response to system failures (if occurs), which can provide partial storage. Even though, the file chunks are replicated and distributed across number of machines, they form a single namespace, so their contents are universally accessible.

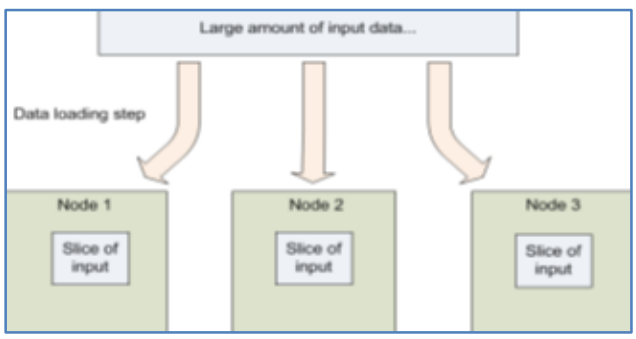


***Figure 2. Hadoop Cluster***

Hadoop limits the amount of communication done by the individual processes as each record is processed by an isolated[11] task, which are different from one another. Programs must be written to a particular programming model, named MapReduce. In MapReduce, records are processed in separately by isolated tasks called Mappers. The output from the Mappers[12] is then moved together into a next set of tasks called Reducers, where results from different mappers can be merged together as shown in Figure 2.

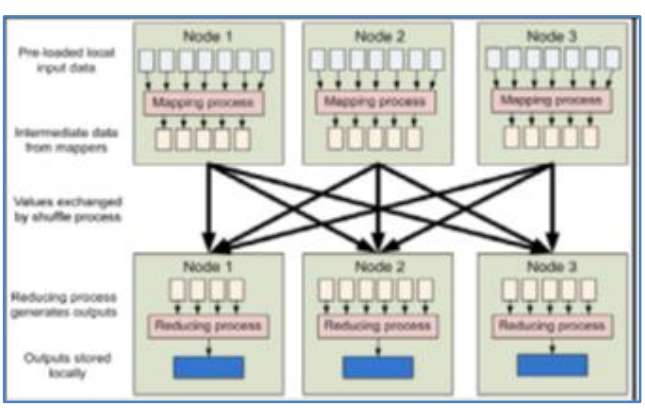


***Figure 3. Mapping of Records Using Hadoop Distributed File System***

**Issues**

Big Data analytics has to face lot issues in the cyber-attacks and it should have detected the intrusions in the network. Dynamic and managed collection,[13] consolidation and correlation of data from any number of diverse data sources, such as network traffic, operating system artefacts and event data (e.g., network devices, IDS). This holistic view of the infrastructure enables defenders to correlate sporadic low-severity events as a result of an ongoing attack.

Anomaly detection based on correlation[14] of recent and historical events; for example, an increased volume of Domain Name System (DNS) traffic from a particular system for a small time period can be due to legitimate user actions. However, if

such a pattern is also identified in historical traffic over a period of days, it is a potential indication of covert data exfiltration. In addition, such correlation can help limit the number of false-positive alerts. Big data analytics solutions increase the quantity and scope of data over, which correlation can be performed.

**CONCLUSION**

The current industry approach, which is focused on real-time detection with emphasis on signature matching, although effective against traditional attacks is unable to address the unique characteristics of APT (advanced persistent search). As mentioned, Big Data analytics currently faces a number of practical limitations and further research is needed for building an operational solution. That said, Big Data analytics will significantly enhance the detection capabilities of defenders, enabling them to detect APT activities that are passing under the radar of traditional security solutions**.**

**Future Enhancement**

Before Big Data analytics can be used in operational environments for the detection of sophisticated threats, a few obstacles need to be overcome. More specifically, there is a need for new detection algorithms capable of processing significant amounts of data from diverse data sources. Additionally, there is a need to further progress issues related to the specific problem of malicious-activity detection using correlated data sources, such as collecting information from untrustworthy sources, storage and processing performance, time synchronisation, meaningful visualisation of information and ensuring the security of sensitive indicators of compromise, among others.

Currently, a small number of proof-of-concept deployments that utilise big data analytics for security event detection exist and show promising results. Research on this promising field needs to be intensified to create robust solutions that can address the multi-dimensional problem of APTs.

**REFERENCES**


1. Cheng W, Min Z, Qiong-mei L. Practices of agile manufacturing enterprise data security and software protection. 2nd International Conference on Industrial Mechatronics and Automation 2010.
2. Chai W. Analyzes and solves the top enterprise network data security issues with the web data mining technology. First International Workshop on Database Technology and Applications 2009.
3. Xuemei L, Yan L, Lixing D. Study on information security of industry management. Asia-Pacific Conference on Information Processing 2009.
4. Oltsik J. Defining the big data security analytics. Network world 1April 2013.
5. Sood AK, Enbody RJ. Targeted cyber-attacks: a superset of advanced persistent threats Security & Privacy. IEEE 2013;11(1):54-61.
6. Hutchins EM, Cloppert MJ, Amin RM. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. 6th International Conference on Information Warfare and Security ICIW 2011. http://www.lockheedmartin.com/content/dam/lockheed/data/corporat e/documents/LM-White-Paper-Intel-Driven-Defense.pdf
7. Apache Hadoop Project. http://hadoop.apache.org/
8. Hadoop Tutorial from Yahoo. Module 7: managing a Hadoop cluster. http://developer.yahoo.com/hadoop/tutorial/module7.html#machines.
9. Shvachko K, Kuang H, Radia S, et al. The Hadoop distributed file system. In: Poc. The 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies MSST 2010.
10. Oltsik J. Strong opportunities and some challenges for big data security analytics in 2014. Network Word 2013. DOI= http://www.esgglobal.com/blogs/strong-opportunities-and-some-challenges-for-big data security analytics-in-2014.
11. Anwar MM, Zafar MF, Ahmed Z. A proposed preventive information security system. IEEE International Conference on Electrical Engineering 2007.
12. MacDonald, N. Information security is becoming a big data analytic problem. Gartner 2012. DOI=http://www.gartner.com/id=1960615
13. Barrett L. Big data analytics: the enterprise's next great security weapon? February 2014.
14. http://www.edupristine.com/courses/big-data-hadoop-program/bigdata-hadoop-course/