iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** II **Month of publication:** February 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Big Data Security Using System Logs

S. Manoj Prabhakar
MAM College of Engineering

Abstract: ***A series of recent high-profile security breaches have underscored that malware prevention strategies are consistently failing to adequately protect enterprises from advanced persistent threats (APTs). It's time to embrace a better alternative—threat detection built on big data analytics. When it comes to prevention methods, information security vendors have traditionally fallen into two camps: either allowing what's on a white list and preventing everything else; or preventing what's on, a blacklist, and allowing everything else. Either way .they're fixated on the tactic of prevention, In addition, the failure of systems, such as firewalls IPS, IDS and Secure Web Gateways, to detect and protect the network is due to the fact that they are policy- and/or signature-based, and can manage only real-time traffic. They are also limited by the capacity of the appliance (CPU, storage, etc.), which means they cannot detect persistent threats. This project proposes and verifies the algorithm to detect the advanced persistent threat early through real-time network monitoring and combinatorial analysis of big data log. Moreover, provide result tested through the analysis in the actual networks of the deduced algorithm.***

Keywords: Advanced Persistent Threats (APTs), Big Data Security Analytics (BDSA), Data Exfiltration

I. INTRODUCTION

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. The challenges include analysis, capture, duration, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on." Scientists regularly encounter limitations due to large data sets which contains complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, RFIDreaders, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5exabytes (2.5×10^{18}) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization. Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration." Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many peta bytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

II. RELATED WORK

A. Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data

Due to rapid development of Internet and technology, all the machines are connected to each other either by networked system or via mobile communication. The users are producing more and more data through communication media in the unstructured form which is highly unmanageable and this management of data is the challenging job. The main focus is to gather the unstructured data from all the terminals, processed the data to convert into structured form so that accessing of the data would be easier. For this, always a track is kept on data, that this data or event belongs to which category. Accordingly, data is analyzed and processed to convert it into meaningful and right information by using the concept of Big Data Analytics. Big Data Analytics accepts the huge

data sets and varied data types, both semi-structured and unstructured like videos, images, audio, web-pages, texts or e-mails etc and convert it into reliable information. Big data analytics describes the simple algorithm for large amount of data without compromising performance. Analysis algorithms are provided directly to database which go beyond the pack and innovate newer more sophisticated statistical analysis. Big Data Analytics use number of tools to do the analysis and processing of data in meaningful way. Hadoop is one of the tools which is aimed to improve the performance of data processing. With the huge amount of processed data available on internet, hackers also become so active with malicious attacks. Hackers target the analyzed data and create threats for information. Big data security analytics is used for the growing practice of organization to gather and analyze security data to detect vulnerabilities and intrusions. The aim is here to make use of Big Data techniques to analyze 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 and targeted attacks have become main subject for government, organization or industry. A subset of threats is Advanced Persistent Threats (APT) which is 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 their target environments.

B. Big Data Analytics for Security Intelligence

The preservation of privacy largely relies on technological limitations on the ability to extract, analyze, and correlate potentially sensitive data sets. However, advances in Big Data analytics provide tools to extract and utilize this data, making violations of privacy easier. As a result, along with developing Big Data tools, it is necessary to create safeguards to prevent abuse. In addition to privacy, data used for analytics may include regulated information or intellectual property. System architects must ensure that the data is protected and used only according to regulations. The scope of this document is on how Big Data can improve information security best practices. CSA is committed to also identifying the best practices in Big Data privacy and increasing awareness of the threat to private information. CSA has specific working groups on Big Data privacy and Data Governance, and we will be producing white papers in these areas with a more detailed analysis of privacy issues. Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for Big Data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses. Off-the-shelf Big Data tools and techniques are now bringing attention to analytics for fraud detection in healthcare, insurance, and other fields.

C. Big security for big data

In the past when the network infrastructure was straightforward and perimeters used to exist, controlling access to data was much simpler. If your secrets rested within the company network, all you had to do to keep the data safe was to make sure you had a strong firewall in place. However, as data became available through the Internet, mobile devices, and the cloud having a firewall was not enough. Companies tried to solve each security problem in a piecemeal manner, tacking on more security devices like patching a hole in the wall. But, because these products did not interoperate, you could not coordinate a defense against hackers. In order to meet the current security problems faced by organizations, a new paradigm shift needs to occur. Businesses need the ability to secure data, collect it, and aggregate into an intelligent format, so that real-time alerting and reporting can take place. The first step is to establish complete visibility so that your data and who accesses the data can be monitored. Next, you need to understand the context, so that you can focus on the valued assets, which are critical to your business. Once the machine data is collected, the data needs to be parsed to derive intelligence from cryptic log messages. Automation and rule-based processing is needed because having a person review logs manually would make the problem of finding an attacker quite difficult since the security analyst would need to manually separate attacks from logs of normal behavior. The solution is to normalize machine logs so that queries can pull context-aware information from log data. For example, HP ArcSight connectors normalize and categorize log data into over 400 meta fields. Logs that have been normalized become more useful because you no longer need an expert on a particular device to interpret the log. By enriching logs with metadata, you can turn strings of text into information that can be indexed and searched.

D. Leveraging Threat Intelligence in Security Monitoring

Malware analysis continues to mature rapidly, getting better and better at understanding exactly what malicious code does to devices. This enables you to define both technical and behavioral indicators to seek within your environment, as Malware Analysis Quant described in gory detail. This is essential because the central strategy of classical AV — file blacklisting — is no longer effective. We need new indicators to detect malware by what it does rather than what it looks like. A number of companies offer information on specific malware samples. You can upload a hash of a malware file: if the recipient has seen it already they will recognize the hash and return the analysis on file; otherwise you upload the whole file for analysis. These services run malware samples through proprietary sandbox environments and other analysis engines to figure out what they do, build detailed profiles, and provide comprehensive reports which include specific behaviors and indicators that can be integrated into monitoring platforms and security controls. These profiles enable you to look for the behavior of malware rather than depending on matching file hashes. The next wave of protection involves looking outside the walls of your own environment to leverage what's happening in the broader world, in order to better prioritize your efforts. The critical limitations of SIEM are the need to know what to look for, and only being able to react after it happens in your environment. Early Warning changes this with external threat intelligence. With a mushrooming variety of threat intelligence sources ready to detail attacks, malware, and tactics seen in the wild; organizations can now look for attacks before they hit, as well as implement preemptive controls to guard against them.

E. Genetic algorithms

The GASSATA system (Genetic Algorithm as an Alternative Tool for Security Audit Trail Analysis) [GASSATA] uses a genetic algorithm to search for the combination of known attacks (expressed as a binary vector, each element indicating the presence of a particular attack) that best matches the observed event stream. A hypothesis vector is evaluated based on the risk associated with the attacks involved, and a quadratic penalty function for mismatched details. In each cycle, the current set of best hypotheses are mutated and retested, so that the probability of false positives and negatives approach zero. This technique, like the neural net approach, offers good performance but does not identify the reason for an attack match. In addition, expressing some forms of behavior, and expressing simultaneous or combined attacks is not possible in this system.

F. Host Log Monitoring

The earliest forms of IDS were batch-oriented systems, periodically searching accumulated system, audit and application logs for signs of suspicious activity [Anderson]. Many modern systems continue to use host logs as a source of raw events. Host logs, comprised of the combination of audit, system and application logs, offer an easily accessible and non-intrusive source of information on the behaviour of a system. In addition, logs generated by high-level entities can often summarise many lower-level events (such as a single HTTP application log entry covering many system calls) in a context-aware fashion. A number of details complicate the use of such logs, however. Foremost among these is the questionable validity of log entries on a victim host, especially those generated after the point of where multiple distributed processes interact. Finally, the quality of information held in logs is frequently low: entries omit critical information, while including large quantities of meaningless detail.

G. Target-based IDS

Another attempt to resolve the ambiguities inherent in protecting multiple platforms lies in combining network knowledge with traffic reconstruction. These target-based ID systems typically use scanning techniques to form an image of what systems exist in the protected network, including such details as host operating system, active services, and possible vulnerabilities. Using this knowledge, a probe can reconstruct network traffic in the same fashion as would be the case on the receiver system, preventing attackers from injecting or obscuring attacks. In addition, this approach allows IDS to automatically differentiate attacks that are a threat to the targeted system, from those that target vulnerabilities not present - thus refining generated alerts (for example, IIS-based attacks on Apache HTTP servers might be ignored). Whether attacks that cannot succeed should be reported is something of a contentious issue - offering a trade-off between lower (and more applicable) security alerts being generated, versus the possibility of recognizing novel attacks when combined with known sequences. In addition, the need to maintain an accurate map of the protected network - including valid points of vulnerability - may reduce the ability of this class of system to recognize novel attacks.

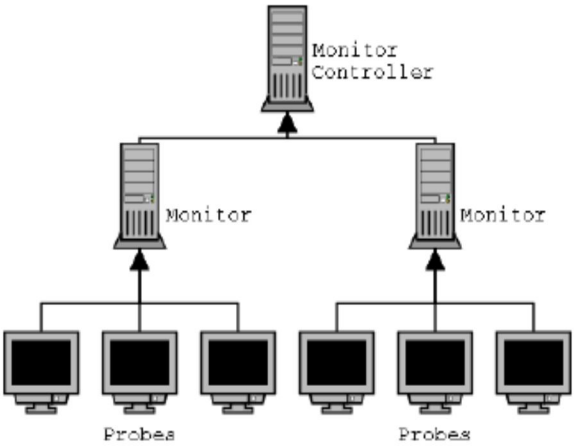
H. Monolithic systems

The simplest model of IDS is a single application, containing probe, monitor, resolver and controller all in one. More advanced monolithic systems use a number of independent probe, monitor and resolver components, each implementing specific techniques. In common between all such systems, however is the fact these focus on a specific host or system - with no correlation of actions

that cross system boundaries. Such systems are conceptually simple, and relatively easy to implement. Their major weakness lies in the ability for an attack to be implemented using a sequence of individually innocuous steps. The alerts generated by such systems may in fact be aggregated centrally - but this architecture offers no synergy between IDS instances.

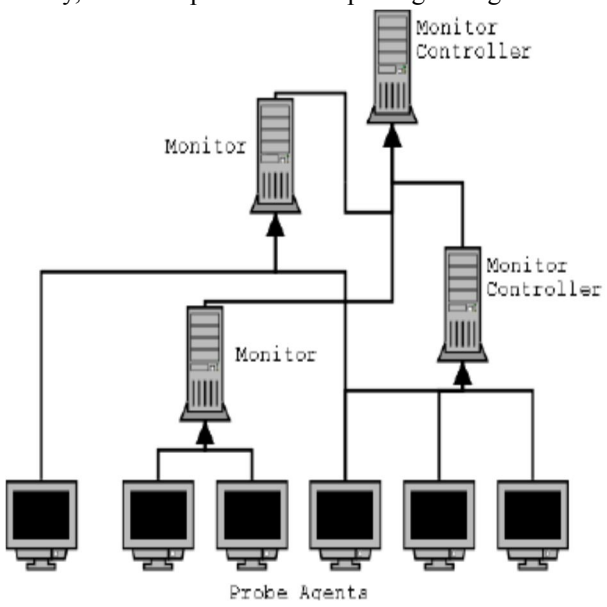
I. Hierarchic systems

If one considers the alerts generated by an IDS instance to be events in themselves, suitable for feeding into a higher-level IDS structure, an intrusion detection hierarchy results. At the root of the hierarchy, lie a resolver unit and controller. Below this lie one or more monitor components, with subsidiary probes distributed across the protected systems. Effectively, the whole hierarchy forms macro-scale IDS. The use of a centralized controller unit allows information from different subsystems to be correlated, potentially identifying transitive or distributed attacks. For example, a simple address range probe, while difficult to detect using a network of monolithic host IDS instances, can be trivial to observe when correlating connections using a hierarchic structure.



J. Agent-based

A more recent model of IDS architecture divides the system in to distinct functional units: probes, monitors, and resolver and controller units. These may be distributed across multiple systems, with each component receiving input from a series of subsidiaries, and reporting to one or more high-level components. Probes report to monitors, who may report to resolver units or higher-level monitors, and so forth. This architecture, implemented in systems such as [AAFID] and [EMERALD], allows great flexibility in the placement and application of individual components. In addition, this architecture offers greater survivability in the face of overload or attack, high extensibility, and multiple levels of reporting throughout the structure.



K. Distributed (GrIDS)

The IDS architectural models described above all consider attacks in terms of events on individual systems. A recent development, typified by the [GrIDS] system, lies in regarding the whole system as a unit. Attacks are modeled as interconnection patterns between systems, with each link representing network activity. The graphs that form can be viewed at different scales, ranging from small systems to the interconnection between large and complex systems (where sub-networks are collapsed into points). This novel approach promises high scalability and the potential to recognize widely distributed attack patterns (such as worm behavior).

L. Honey nets

The concept of a honey trap is simple: a committed sacrificial system placed in a striking or ubiquitous position on a network and designed to receive attacks which contain a good example of such a system, manually implemented. Originally, honey traps consisted of heavily monitored, real systems or virtual systems implemented by software. A recent innovation was the use of a so-called honey neto: an entire network of systems, in its entirety sacrificial. Separating this network from the outside world is a firewall - configured to allow unrestricted incoming access, but limit outgoing access. In this manner, an attacker is prevented from using the machines in the honey net as a relay point for attacking other systems. Any traffic to or from the honey net is fully logged. Experimentation with such systems has provided fascinating details, ranging from the development of passive fingerprinting techniques, to insights into the social interaction of hackers [Honeynet].

M. APT prevention in BDST system

Attackers today are motivated to steal intellectual property and financial data. While there are well-publicized nuisance attacks by organized groups with Twitter accounts, the largest threats are from state-sponsored groups and organized crime. Known as advanced persistent threats (APTs), these groups target organizations with a specific goal in mind. APT groups are well educated, well funded and highly motivated. They use a number of techniques ranging from sophisticated hacking to social engineering in order to bypass traditional security tools and gain access. Simple techniques such as targeted phishing (known as spear phishing) that looks like legitimate corporate email to infect organizations are very common with state-sponsored groups. Other organized crime groups in both the US and Eastern Europe largely use malware. Once inside, APTs begin executing reconnaissance activities using methods that typically go unnoticed. A recent Verizon report noted that the vast majority of breaches went undetected for months and were only discovered after the theft was reported. Clearly, the new breed of APTs has the capabilities to defeat traditional security solutions, but also navigate inside the network and execute exfiltration activities with impunity. This project will collect DHCP server logs in addition to domain controller Kerberos events. Assume that domain controllers and servers that routinely have different users logging on to them have static IP addresses and do not appear in the DHCP logs. Now, we begin building the core of this detection scenario with an active list called Workstation Current IP. This list comprises two columns: computer name and IP address. The list builds itself from lease and lease renewal events that are collected from DHCP servers. The list maintains one unique row for each IP address and keeps the computer name found in the most recent event for that IP address. Thus, we have a list of each workstation and its current IP address, automatically updated when a workstation receives a new address. Any computers that multiple users share should be filtered from this list by naming convention or by regularly extracting a list of such computers from Active Directory, using appropriate criteria.

III. CONCLUSION

The goal of massive knowledge analytics for security is to get unjust intelligence in real time. Though massive knowledge analytics have important promise, there area unit variety of challenges that has got to be overcome to comprehend its true potential. The subsequent area unit just some of the queries that require to be addressed:

A. Knowledge provenance

Credibility and integrity of knowledge used for analytics. As massive knowledge expands the sources of knowledge it will use, the trustiness of every knowledge supply must be verified and therefore the inclusion of ideas like adversarial machine learning should be explored so as to spot maliciously inserted knowledge.

B. Privacy

We'd like restrictive incentives and technical mechanisms to attenuate the quantity of inferences that massive knowledge users will

build. CSA includes a cluster dedicated to privacy in massive knowledge and has liaisons with NIST's massive knowledge unit on security and privacy. We have a tendency to decide to turn out new pointers and white papers exploring the technical means that and therefore the best principles for minimizing privacy invasions arising from massive knowledge analytics.

C. Securing massive knowledge stores

This document targeted on mistreatment massive knowledge for security, however the opposite facet of the coin is that the security of massive knowledge. CSA has created documents on security in Cloud Computing and conjointly has operating teams specializing in distinctive the simplest practices for securing massive knowledge.

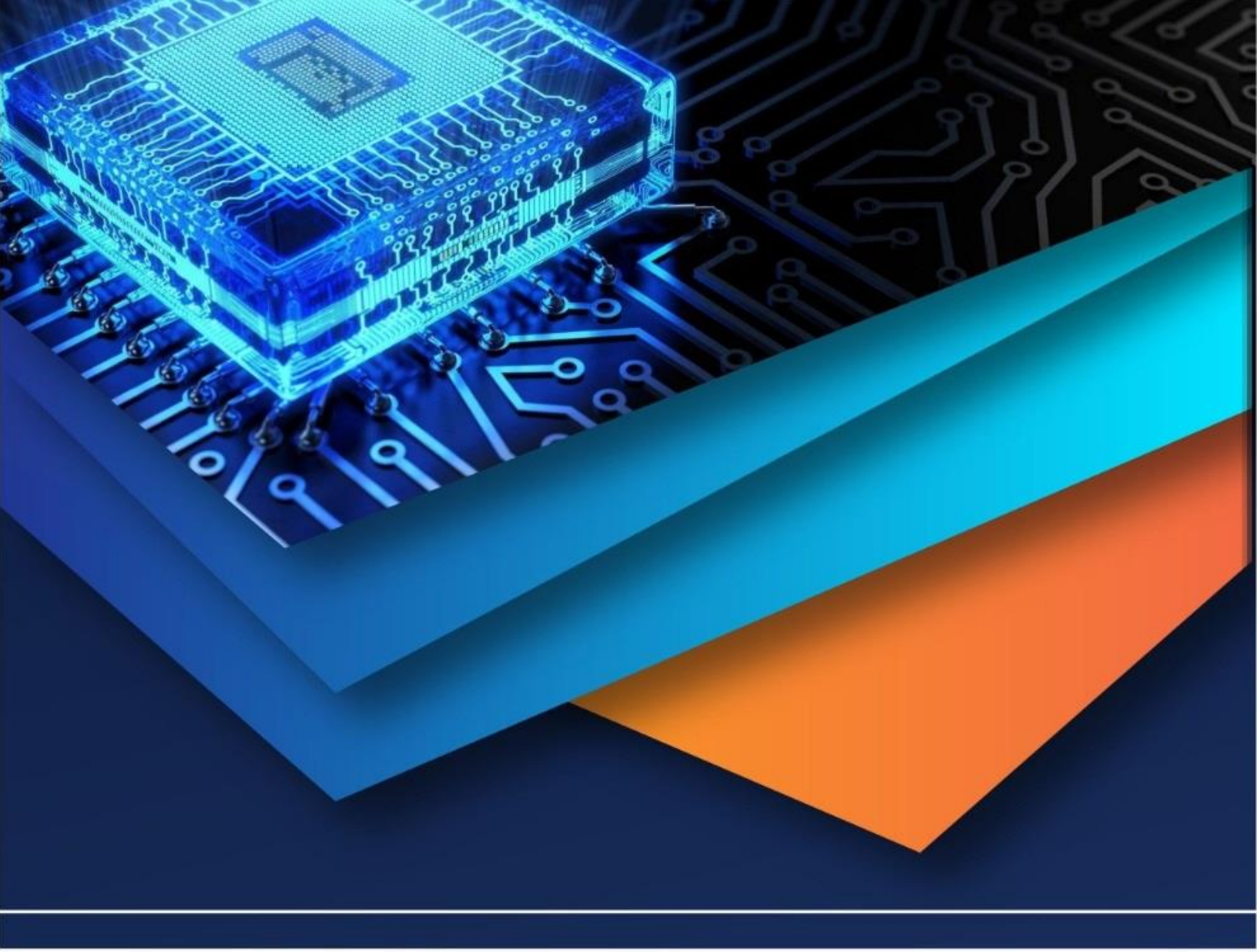
D. Human-computer interaction

Massive knowledge may facilitate the analysis of numerous sources of knowledge; however somebody's analyst still should interpret any result. Compared to the technical mechanisms developed for economical computation and storage, the human-computer interaction with massive knowledge has received less attention and this is often a region that must grow. an honest beginning during this direction is that the use of mental image tools to assist analysts perceive the information of their systems.

We hope that this primary report on massive knowledge security analytics outlines a number of the elemental variations from ancient analytics and highlights attainable analysis directions in massive knowledge security.

REFERENCES

[1] Wang Cheng, Zeng Min, Liu qiong-mei. Practices of Agile Manufacturing Enterprise Data Security and Software protection. 2nd International Conference on Industrial Mechatronics and
Automation, 2010.
[2] Wenguang Chai. Analyzes and solves the Top Enterprise Network Data Security Issues with the Web Data Mining Technology. 2009 First International Workshop on Database Technology and Applications, 2009.
[3] Li Xuemei, Li Yan2, Ding Lixing. Study on Information Security of Industry Management. Asia-Pacific Conference on Information Processing, 2009.
[4] J. Oltsik. Defining the big data security analytics. Networkworld, 1 April 2013.
[5] A. K. Sood, R.J. Enbody "Targeted cyber attacks: A Superset of advanced persistent threats' Security & Privacy, IEEE Volume 11, Issue 1, pages 54-61, Jan-feb. 2013
[6] Eric M. Hutchins, Michael J. Cloppert , Rohan M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains" ,6th
International Conference on Information Warfare and Security(ICIW2011) http://www.lockheedmartin.com/content/dam/lockheed/data/corporate/documents/LM-White Paper-Intel-Driven-Defense.pdf
[7] A.K.Sood, R.J. Enbody "Targeted Cyber attack: A superset of advanced persistent threats" Security & Privacy, IEEE Volume 11 Issue 1, pages 54-61, Jan-Feb, 2013.
[8] Apache Hadoop Project http://hadoop.apache.org/
[9] "Hadoop Tutorial from Yahoo!", Module 7: Managing a Hadoop Cluster http://developer.yahoo.com/hadoop/tutorial/module7.html#machines
[10] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop distributed file system", in poc. The 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY