IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Applicable to BIG DATA IS BIAS-BASED ANALYTICS FOR SECURITY INTELLIGENCE

**[1]ANN MARY P J, [2]ANILA JOSE**

**[1]Msc Scholar, [2]Assistant Professor**

[1,2]Department of Computer Science

[1,2]St.Joseph's College (Autonomous), Irinjalakuda, Thrissur, India

**ABSTRACT:** Cloud Technology, Data Mining, Hadoop, and MapReduce are some of the technologies that are used to analyze Big Data today in the world of Information Technology. For several decades, securing valuable data from intruders, viruses, and worms has been a challenge. To preserve the data, several technologies and procedures have been developed. In this study, we analyzed numerous security and privacy approaches proposed by various researchers, as well as the benefits and drawbacks of each methodology. The term Big Data refers to technologies for managing and analyzing large amounts of information that transcend the capabilities of traditional data processing techniques. The three characteristics of Big Data differentiate it from traditional technologies: volume of data (volume), speed of data transmission (velocity), and variety of structured and unstructured data types (variety). The study examines how Big Data has aided in the advancement of security analytics by providing new tools and opportunities for analyzing large amounts of structured and unstructured data.

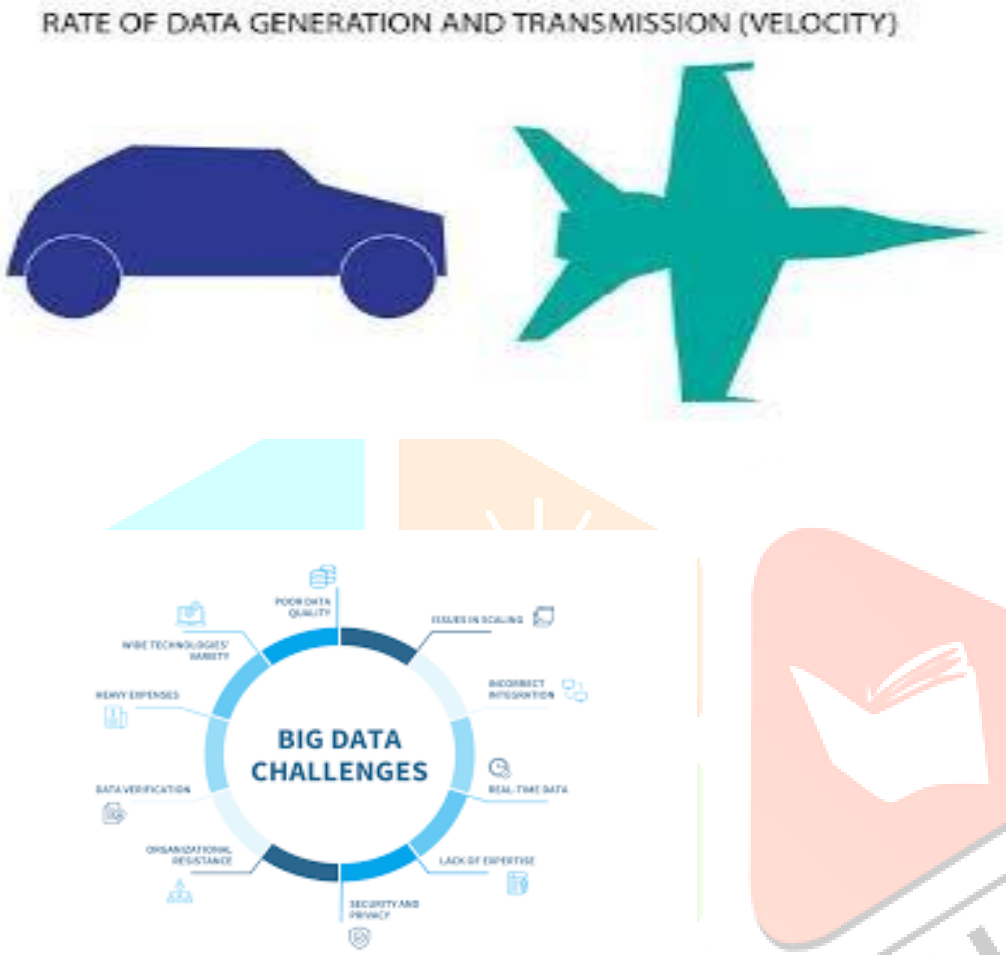
**KEYWORDS:** Data analysis, cloud computing, data mining, Hadoop, MapReduce, and security and privacy methodologies.

**INTRODUCTION:** Cyber-attacks by hackers have escalated considerably in recent years, according to leading security organizations such as Symantec and McAfee. Businesses are losing trillions of dollars as a result of cyber assaults, in addition to millions of personal and financial records being lost. Artificial Neural Networks, Support Vector Machines and other techniques for machine learning are examples.

Users continue to be concerned about data security as the number of data increases. Since troglodytes used guard dogs to protect their belongings which later evolved into moats around castles, and then we had security guards for individual houses, and later home security systems, and then cryptography was developed. The BI security of a corporation should be built on three primary best practices or difficulties. The BI security of a corporation should be built on three primary best practices or difficulties. The first difficulty is that incoming data could be intercepted or corrupted while in transit. Data stored on the cloud or on-premise systems are equally vulnerable to theft or enslavement. Last but not least is the output data, which is seemingly unimportant but which can be a gateway for hackers or other malicious parties.

Big Data security includes four different components:

(1) Infrastructure security,
(2) Data privacy,
(3) Data management, and
(4) Integrity and reactive security.

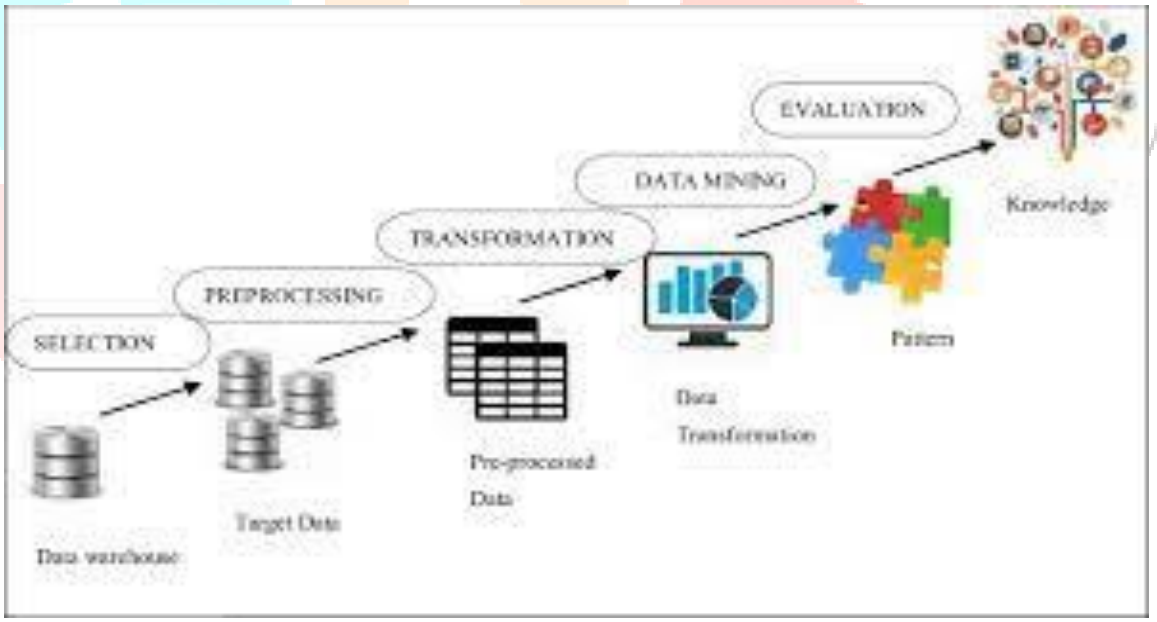



The following ways in which big data analysis can transform security analytics are:

(a) It can aggregate data from various sources within organizations as well as from external sources to form something known as a vulnerability database by combining the required data.
(b) The following ways in which big data analysis can transform security analytics are: (a) It can aggregate data from various sources within organizations as well as from external sources to form something known as a vulnerability database by combining the required data.
(c) All related information is presented as a single view.
(d) The system analyzes streaming data in real-time, using previous results as feedback to the system.

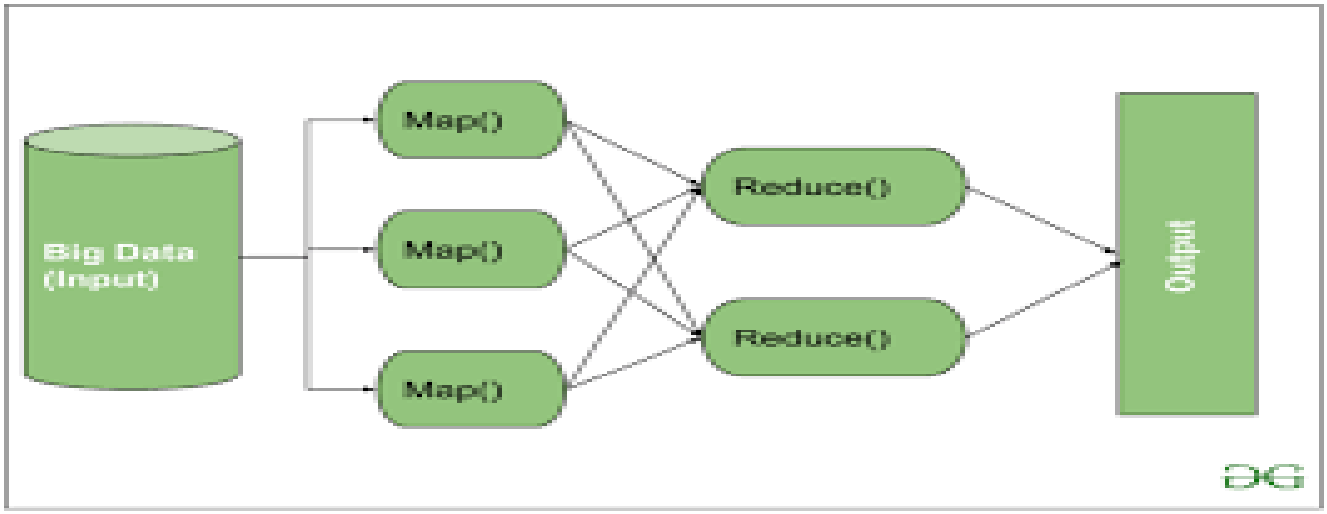
Cloud computing and big data are increasingly considered to be complementary technologies. Big data enables users to conduct distributed queries over various datasets using commodity computing and provide resultant sets in a timely way. Hadoop, a type of distributed data-processing platform, offers the fundamental engine in cloud computing.

Big data allows users to do distributed queries over multiple datasets utilizing commodity computing and receive the results promptly. It primarily addresses the security challenges that arise when applying data mining techniques on a large scale, as well as a discussion of several processes that can aid in data security. The primary idea is to identify distinct categories of consumers who are concerned about data mining application security Recent PPDM research has focused on ways to reduce the security risk posed by data mining tools. In this sense. In this sense, "information gathering" is viewed as a synonym for the term "Knowledge Discovery from Data" (KDD), which emphasizes the mining procedure's goal




Big data uses Hadoop, a free Java-based programming framework, to process massive amounts of data in a distributed computing environment. Hadoop uses a Master/Slave model to process massive volumes of data over a cluster of computers, and applications can run on systems with thousands of nodes and terabytes of data. In the event of a failure, taking care of the network uses a Distributed File System (DFS), which enables fast data transfer speeds and large-scale data storage. Because the system can continue to function regularly, the chances of a system failure are reduced. Hadoop is a scalability-oriented computational system. Google, Yahoo, and other well-known companies use the Hadoop Framework because it is scalable, low-cost,

customizable, and fault-tolerant.




**WORKS RELATED TO BIG DATA:** A lot of work has been done in the field of big data.

Several frameworks have arisen as a result of the growth of big data technologies. It has been suggested that massive amounts of data from various sources be analyzed. The majority of these frameworks employ the Hadoop framework. HDFS is a data-storage distributed file system.

Tazaki et al. demonstrated MATATABI [9], a security monitoring platform that conducts multi-layer threat analysis on a range of data sources, such as network device traffic and human interaction logs. When it comes to data analysis, MATATABI uses Hive and employs a SQL approach. An initial query looks for suspicious indicators of threat in the collected datasets, such as an IP address belonging to a botnet, and then tries to find the same signals in other datasets to find correlations between them. Finally, particular features from the correlated indicators are saved in a blacklist, such as IP addresses, timestamps, and so on.

blacklist.

Marchal et al. provided an architecture for security monitoring in a local enterprise network context. The architecture receives data from various sources and distributes it. Data sources include honeypots, DNS traffic, HTTP traffic, and IP-Flow records from edge routing devices. Instead of the SQL-like queries used by MATABI. In this technique, correlation is achieved by computing scores based on the data collected. The security administrator can set thresholds that indicate whether the system should create a warning, allow traffic, or reject traffic in response to a specific risk score. However, this study is only appropriate for single domain analysis, such as network intrusion detection and prevention, and it currently lacks real-time intake and processing.

Tan et al. employed a data correlation strategy to create a platform for data sharing across existing Intrusion Detection Systems (IPSs) and Intrusion Prevention Systems (IPSs) (IDSs). CIDs (collaborative intrusion detection systems) share network information with IDSs deployed at network edges. They also correlate all suspicious data gathered from several IDSs to improve the efficacy of intrusion detection.

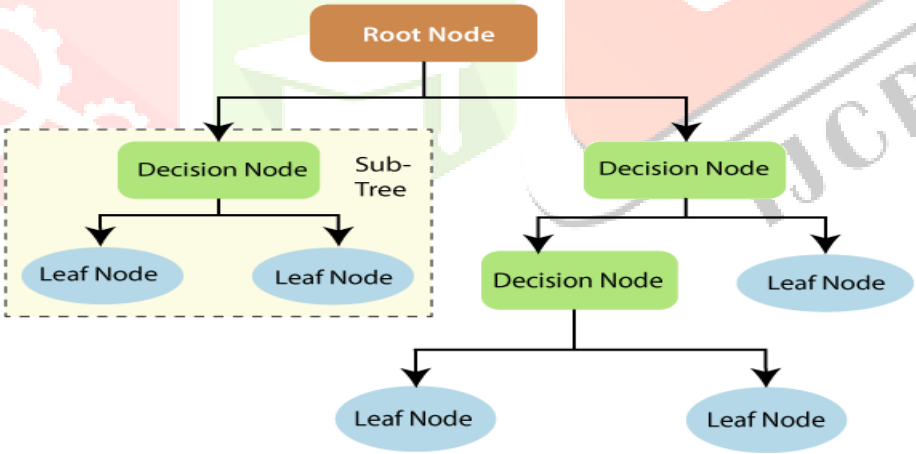
Other approaches, such as employing meta-models and statistical data models to detect network anomalies on a per-flow basis, involve complex machine learning algorithms. According to Yang et al., data ingestion is more efficient and scalable when translated into a meta-model since raw data only needs to be retrieved once, and a meta-model is substantially less in size, allowing it to be resident in system RAM for faster access.

**WHY BIG DATA ANALYTICS FOR SECURITY:** Large amounts of data from several sources must be collected for security monitoring to be holistic and comprehensive. The issue of storing collected datasets and properly executing computations on multidimensional data (a mix of structured and unstructured data) becomes more obvious as data volume grows and data sources become more diverse.

One of the main reasons why big data can be used with security is that an organization needs to maintain its traditional security information for a longer period to analyze the data. Historical study of this data has the perspective of uncovering unacknowledged security vulnerabilities and identifying breaches in security over some time. This is possible using big data analytics as the storage cost has decreased tremendously and while the traditional data warehouses retain data for a specific period, big data can go about maintaining a piece of data indefinitely. Hence allowing us to find patterns in data so as toto possible security attacks.

Second, data sources that aren't often used for security can assist a company in determining its assets and liabilities. aspects that must be safeguarded and observed traditional security solutions are mostly used to protect structured data. Big data analytics deals with these issues since they are inflexible by nature. And combine organized and unstructured data, obviating the requirement for a separate database can therefore carry out any predetermined data forms. examination of data gleaned from sources such as email, Web material, social media, and business papers.

Third, additional analytical procedures must be done to extract security-related insights from huge data sets, necessitating more processing time. Big data solutions such as Hadoop and NoSQL databases have improved the speed with which complex queries are processed and unique patterns within massive blocks of data can be identified. The asynchronous analysis must be carried out in real-time, and once completed, the results must be communicated back to the real-time element for the real-time element to deliver an optimal solution over time.




The difference between the DT algorithm and RF algorithm is that the establishing of the root nodes and the segregating nodes are done randomly in the latter. The RF follows the bagging method to generate the required prediction. The ultimate output is produced by the leaf node of each DT.. In the Random Forest classifier, the selection of the final output follows the majority-voting system. That is the output chosen by the majority of DT becomes the final output of the RF system.

**THE GREAT CHALLENGE OF BIG DATA SECURITY**: Organizations employed several ways of deidentification to maintain security and privacy, according to Minit Arora and Dr. Himanshu Bahuguna(2016)[4]. Oral and written commitments are the most usual way to protect confidentiality and privacy. This strategy, however, has been proven to be incorrect throughout history. When sharing and aggregating data across dynamic, dispersed data systems, passwords, controlled access, and two-factor authentication are low-level, but often utilized technical solutions to guarantee security and privacy. Both the purposeful sharing of rights and the continuance of permissions after they are no longer required or permitted can potentially break access permissions like these. Cryptography is a more advanced technological answer. AES and RSA are two well-known encryption techniques. According to recent reports, the National Security Agency (NSA). Because it takes specialized employees to read and interpret the findings, and the software might be exploited to monitor individual behavior rather than securing data, this technology is difficult and costly to adopt on a wide scale or with distributed data systems and users. As a result, traditional de-identification approaches are no longer effective in the era of Big Data due to the widespread use of de-identification techniques. As the amount of data grows, protecting Big Data security and privacy becomes more complex. Even anonymous data may often be re-identified and attributed to specific individuals, according to computer scientists.

**SECURITY INTELLIGENCE TECHNOLOGIES:-** We need effective threat detection and investigation services to fight strongly against every cybercrime. To terminate the problem in the nib, an ideal security team requires a system that is flexible enough to gather data from many security specialized domains, which may unleash potential security dangers and risks and warn analysts in seconds. The following are some models that have a lot of features like an integrated security system:-

A. RSA Security Analytics: RSA security analytics is security software developed by EMC2. It has transformed security management from a traditional log-centric approach to one with exceptional workflow ethics, and thus stands out among other solutions. The RSA security analytics solution provides:

1) Complete infrastructure visibility: This can collect all security-related data from a variety of sources. Analysts may now access data related to security, risks, and malicious behaviors in a single, consolidated format. Real-time analytics and historical data queries are also available through the architecture. As a result, this architecture enables real-time threat modification, inquiry analysis, as well as historical data retention and archiving.

2) Agile Analytics: This security platform gives analysts with rapid investigative tools as well as easy tools for rapid analysis that aid in the deep study of data collection, allowing for better decision-making. Depending on the end-users connected to the infrastructure, this uses signature-free analytics to detect malicious users and actions. It also offers the ability to replay and reproduce events exactly as they occurred.

3) Actionable Intelligence: RSA's threat intelligence assists analysts in obtaining optimal solutions by implementing current threat feeds. The intelligent system embedded into the architecture is constrained by rules, reports, and watch lists. This allows analysts to acquire a better understanding of a problem and prioritize responses or countermeasures accordingly.

4) Streamlined Readiness and Response: The platform offers functions that allow security teams to access and monitor risks by streamlining activities connected to preparedness and response through workflow systems that explicitly define and activate responses if threats are realized, as well as tools to access and monitor the presently open processes.


B. The Beehive Model of Actionable Intelligence

Advanced Persistent Threats are attacks on high-value intellectual property or physical systems (APT).

Unlike other types of malware that spread quickly, such as Trojans, viruses, and worms, APT attackers are small and slow. Low mode refers to keeping a low profile in networks, while slow mode refers to staying in a system for an extended period.

## CONCLUSION:-

The main purpose of big data analytics is to gather actionable intelligence in real-time. However, to demonstrate its entire potential, big data analytics must overcome several obstacles and keep the big commitment that it made in the past in the field of security some issues must be addressed.to make Big Data analytics a household brand in the security industry are:

(1) Data Provenance: Big Data is a collection of information. information obtained from a variety of sources, and Before being used, it must be checked for authenticity and integrity. Any analytics is applied to it because certain things happen. data is injected in a nefarious manner.

(2) Privacy: To give a high degree of data security, Big Data analytics may infringe on user privacy. As a result, correct standards must be established to provide an ethical yet effective security system with Big Data analytics.

(3) Securing Big Data Stores: Using Big Data with security is only one side of the coin; the main issue to deal with is big data security; it is critical to secure massive data collections.

(4) Human-Computer Interaction: Big Data analyses a variety of data sources, but the results must still be interpreted by a human analyst.

# REFERENCE:-

[1] "Big Data Analytics for Security," by Alvaro A. Cardenas, Pratyusha K. Manadhata, and Sreeranga P. Rajan, IEEE Security and Privacy magazine, Feb 2015.

[2] "Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks," by Ting-Fang Yen, Alina Oprea, Kaan Onarlioglu, Todd Leetham, William Robertson, Ari Juels, and Engin Kirda.

[3] http://darkreading.com/case-study-in-big-data-analytics

[4] Big Data Analytics for Security Intelligence by the R Cloud Security Alliance

[5] "Big Data Fuels Intelligence-Driven Security," by Sam Curry, Engin Kirda, Addie Shwartz, William H. Stwart, and Amit Yoran. January 2013

[6] IBM Big Data Security Intelligence

http://www-03.ibm.com/security/solution/intelligence-big-data/

[7] Intelligence and analytics in security http://www03.ibm.com/software/products/en/category/security-intelligence

[8] An interview with Ben Wuest about the use case for big data and security analytics http://securityintelligence.com/the-use-case-for-bigdata-and-security-an-interview-with-ben-wuest-about-the-use-case-for-bigdata-and-security-an-interview-with-ben-wuest-about-the-use-case-for-bigdata-and-security-analytics-an-interview-with-b

[9] Four Different Types Of Big Data Analytics And An Example Of Their Application

http://www.ingrammicroadvisor.com/data-center/four-types-of-bigdata-analytics-and-examples-of-their-use

[10] IBM Information Management and Big Data http://www01.ibm.com/software/data/bigdata/enterprise.html

[11] IBM, "Security Intelligence Extending with Big Data Solutions."

January of this year.

https://www.emc.com/collateral/datasheet/security-analytics-overview-ds.pdf [12] RSA Security Analytics

[13] "What You Need to Know About Security Intelligence with Big Data," by Vijay Dheap, July 2013.

https://technet.microsoft.com/enus/library/cc959354.aspx [14] Types of Data Attacks

[15] RSA Security Analytics in action.

http://www.emc.com/about/news/press/2013/20130130-01.htm.