Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security

Danda B. Rawat, *Senior Member, IEEE*, Ronald Doku, and Moses Garuba

Abstract—“Knowledge is power” is an old adage that has been found to be true in today’s information age. Knowledge is derived from having access to information. The ability to gather information from large volumes of data has become an issue of relative importance. Big Data Analytics (BDA) is the term coined by researchers to describe the art of processing, storing and gathering large amounts of data for future examination. Data is being produced at an alarming rate. The rapid growth of the Internet, Internet of Things (IoT) and other technological advances are the main culprits behind this sustained growth. The data generated is a reflection of the environment it is produced out of, thus we can use the data we get out of systems to figure out the inner workings of that system. This has become an important feature in cybersecurity where the goal is to protect assets. Furthermore, the growing value of data has made big data a high value target. In this paper, we explore recent research works in cybersecurity in relation to big data. We highlight how big data is protected and how big data can also be used as a tool for cybersecurity. We summarize recent works in the form of tables and have presented trends, open research challenges and problems. With this paper, readers can have a more thorough understanding of cybersecurity in the big data era, as well as research trends and open challenges in this active research area.

Index Terms—Big data security, big data driven security, IDS/IPS, data analytics

1 Introduction and Background

Over the past 15 years, data has increased exponentially in various applications which has led to the big data era (Fig. 1). It is worth noting that big data has some peculiar features which can be leveraged for various purposes (Fig. 2). One of these is the use of big data for detecting risks or attacks. “As our technological powers increase, the side effects and potential hazards also escalate” is a quote by Alvin Toffler which perfectly sums up the world we live in now. Hacking was at first akin to public defacements of things. Hackers hacked for fun and for notoriety. However, these days, attacks are more calculated and motivated. Nations are accusing each other of hacking. There is also a significant rise in industrial espionage which can either be from nation-state or competing entities trying to gather information or to take away a competitor’s edge as to increase their own. Additionally, we are seeing this across industries from health care to retail to government to education to the financial sector. Thus, with this much susceptibility and hacking advancements, cybersecurity has become an important field in computer science. Cybersecurity aims at reducing the attack vectors/points to a minimal, because it is impossible secure every attack point. An attacker only has to be successful once which has consequently made the job of securing systems very challenging. The number of attackers out there out-number the people trying to protect it. This is because there is so much information out there that can turn anyone into an attacker. With this in mind, cybersecurity has now gone beyond the traditional way of only focusing on prevention to a more sophisticated PDR paradigm which is: Prevent, Detect and Respond (PDR). Big data is expected to play a major role in this emerging PDR paradigm.

Big data is now a common slogan used to mean the generation of large volumes of data. Enormous amount of data are being generated at an alarming rate. This is due to the growth of the Internet. Laney [1] came up with the term the three V’s which he associated with big data. These terms were volume, velocity, and variety. In addition to 3 V’s, there is fourth V which is veracity. Volume represents the fact that the data being generated is enormous, velocity represents the fact that data is being generated at an alarming rate, and variety represents the fact that the data being generated comes in all types of forms. Big Data could be explained simply as data at rest according to Miloslavskaya et al. [2]. They also highlighted the difference between big data, data lake, and fast data. Data lake holds a large amount of raw data in its original format. Fast data can be time sensitive data which may either be structured or unstructured, which is usually acted upon right away.

We have more and more data coming, and they are moving from terabytes to petabytes, which are becoming unfamiliar realms [3]. Thus, we need to find new ways of accommodating this data, and there is the need to develop models and algorithms that will enable us to work on these data, to gain insights from it. This is where Big Data Analytics (BDA) comes in. This paper explores research work

- *Authors are with the Data Science and Cybersecurity Center (DSC^2), Department of Electrical Engineering and Computer Science, Howard University, Washington, DC 20059 USA.*
E-mail: db.rawat@ieee.org, rdoku@bison.howard.edu, mgaruba@howard.edu.

Manuscript received 4 Apr. 2018; revised 19 Jan. 2019; accepted 19 Mar. 2019. Date of publication 25 Mar. 2019; date of current version 9 Dec. 2021.
(Corresponding authors: Danda B. Rawat)
Digital Object Identifier no. 10.1109/TSC.2019.2907247

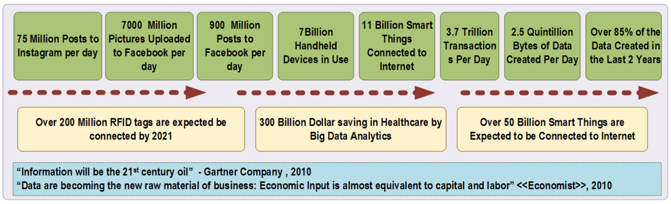
75 Million Posts to Instagram per day
7000 Million Pictures Uploaded to Facebook per day
900 Million Posts to Facebook per day
7Billion Handheld Devices in Use
11 Billion Smart Things Connected to Internet
3.7 Trillion Transaction s Per Day
2.5 Quintillion Bytes of Data Created Per Day
Over 85% of the Data Created in the Last 2 Years
Over 200 Million RFID tags are expected be connected by 2021
300 Billion Dollar saving in Healthcare by Big Data Analytics
Over 50 Billion Smart Things are Expected to be Connected to Internet
"Information will be the 21st century oil" - Gartner Company , 2010
"Data are becoming the new raw material of business: Economic Input is almost equivalent to capital and labor" <<Economist>>, 2010

Fig. 1. Big data is increasing exponentially making security harder.

done on big data enabled security and securing big data (which are categorically presented in Fig. 3).

Although there are related survey papers [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] on big data security (further details, please refer to Section 4), we present more up to date approaches, insights, perspectives and recent trends on the rapidly advancing research field of big data in the cybersecurity domain. Our approach to this covers the research work done on how big data is used as a security tool and the emergence of big data as high value asset resulting in research work done on how to secure big data. Specifically, the main contributions of this paper include:

- Presenting a comprehensive study on security aspects of big data by categorizing it into two parts: security using big data and big data driven security.
- Presenting a summary of attacks and countermeasures for big data in a tabular form for a side-by-side comparison.
- Presenting a discussion of research challenges, recent trends, insights and open problems for big data in cybersecurity.

The remainder of this paper is organized as follows. We first classify our work into two major sections (Sections 2 and 3). We provide a comprehensive study of security using big data as well as securing big data. For each category, we present the related recent state-of-the-art literature for the different approaches. Section 2 focuses on the use of big data as a security mechanism. Section 3 tackles how big data is being protected. Section 4 presents relevant survey papers along the line of this paper and the distinction of this paper from the rest of the surveys. Section 5 presents some research challenges and future directions in this area. Finally, we summarize the paper in Section 6.

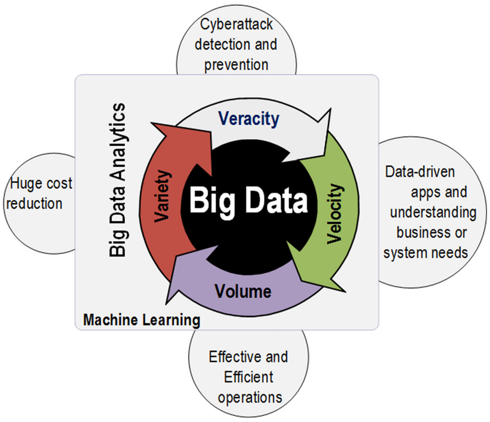
Cyberattack detection and prevention
Huge cost reduction
Data-driven apps and understanding business or system needs
Effective and Efficient operations
Big Data Analytics
Machine Learning
Veracity
Variety
Velocity
Volume
Big Data

Fig. 2. Big data offers typical benefits to business such as informed decisions, competitive advantages and data-driven cybersecurity.

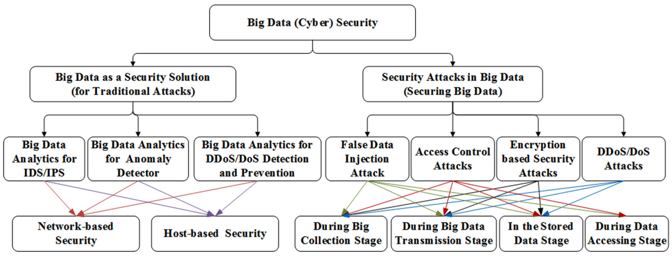
Big Data (Cyber) Security
Big Data as a Security Solution (for Traditional Attacks)
Security Attacks in Big Data (Securing Big Data)
Big Data Analytics for IDS/IPS
Big Data Analytics for Anomaly Detector
Big Data Analytics for DDoS/DoS Detection and Prevention
False Data Injection Attack
Access Control Attacks
Encryption based Security Attacks
DDoS/DoS Attacks
Network-based Security
Host-based Security
During Big Collection Stage
During Big Data Transmission Stage
In the Stored Data Stage
During Data Accessing Stage

Fig. 3. Big data (analytics) as a security solution and security attacks that are unique to big data in a typical big data enabled systems.

2 Security Using Big Data

Top security companies joined forces to share information with each other in an attempt to gather intelligence from the shared data (SecIntel Exchange). Their goal was to provide reliable security tools for their clients, and to achieve that, they had to learn as much as possible from evolving threats that were developed each day. They understood the power of collaboration for the greater good. This was needed because with the rise of polymorphic malware and other evolving threats, they needed a lot information on these threats in order to fully understand what they were dealing with and how to counteract against it. The traditional approaches of classifying malware were proving to be futile. SecIntel Exchange data provided them with the opportunity to derive actionable insights from voluminous data. Human analysis and traditional methods such as database storage could however not keep up with the pace of the data that was being generated [17]. There was the need to adopt modern approaches. As seen in a case study conducted by Zions Bancorporation [18], it would take their traditional Security Information and Event Management (SIEM) systems between 20 minutes to an hour to query a month's worth of security data. However, when using tools with Hadoop technology, it would only take about one minute to achieve the same results. As such BDA has become an important tool in cybersecurity. Several studies have shown that the traditional approaches and human analysts can not keep up with the big data. BDA is one of the best solutions to combat these issues.

2.1 Big Data Analytics (BDA) as a Tool to Combat Diverse Attacks

Aypical attacks that can be subdued using big data analytics are depicted in Fig. 4). "If we know the enemy and ourselves, we need not fear the result of a hundred battles" is an excerpt from the Art of War written by the famous Chinese general, Sun Tzu. In other words, it may not be possible to know

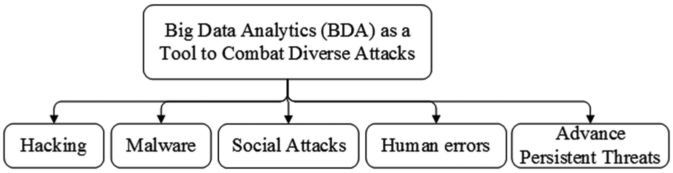
Big Data Analytics (BDA) as a Tool to Combat Diverse Attacks
Hacking
Malware
Social Attacks
Human errors
Advance Persistent Threats

Fig. 4. Big data (analytics) can combat diverse attacks.

enough about our enemy, but it is definitely possible to know all that we can about ourselves and the assets we protect. To do that, we have to gather facts about the asset. This is made possible by the data it generates. This data needs to be analyzed and insights need to be drawn. BDA can help prepare, clean, and query heterogeneous data with incomplete and/or noisy records [19], something that would be hard for humans to do. Analyzing data tends to be hard when the data is heterogeneous as [20] discovered. In their work, they presented a platform targeted at achieving real time detection and visualization of cyber threats which they called OwlSight. The platform had several building blocks (data sources, big data analytics, web services and visualization) and had the ability to collect large amounts of information from a variety of sources, analyze the data and output the findings on insightful dashboards. They did face some issues with the heterogeneity of the data. However, for machines to do the work effectively, they need to have some form of human element. Understanding a problem is half the problem solved. The authors in [21] understood this and addressed this issue by coming up with an approach that merged big data analytics with semantic methods with the aim of trying to gain further insights on the heterogeneous data by understanding it semantically. BDA can be used to gather insights making it an essential tool in cybersecurity. However, the features of big data (four V's) also make deriving insights a hard task to accomplish.

In the 2017 Data Breach Investigations Report done by Verizon, it was reported that attacks tend to come from different sources. 62 percent of the attacks involved hacking, 51 percent used malware and 43 percent were social attacks. 14 percent were a result of human errors. As such, the attacker sometimes relies on human factor in order to execute a successful attack. In such scenarios, people instead of technology become the target of an attack. Email scams and phishing are the most common form of these attacks. In a recent study [22], 52 percent of successful email attacks get their victims to click within an hour and 30 percent within 10 minutes. The authors in [23] looked into the role of big data in such attack scenarios. To gain further insights, the authors conducted two studies. The first study involved the Enron email dataset. The second study was carried out on undergraduate students to observe how email phishing broke security systems based on user behaviours. The collected data was then analyzed using Enronic software which was followed by the categorization of email topics. The authors found that, phishers or attackers can understand the behavior of email users using big data analytics, and therefore are able to generate phishing emails that created security threats based on the insights they gathered. The authors planned on proposing a framework for addressing security threat in email communication in the future. In another work, a big data enabled framework was proposed in [24] with the aim of defending against spam and phishing emails by using a global honeynet. Their framework collected data from different sources such as pcap files, logs from a honeynet, black listed sites and social networks for analysis. The framework used Hadoop and Spark for the processing of the collected heterogeneous data which was stored in Hadoop Distributed File System (HDFS). However, this framework does not provide real-time analysis for big data.

Another form of attack is Advanced Persistent Threats (APT) which are sophisticated, well-planned attacks [25]. APTs are very hard to detect, and the challenge of detecting and preventing advanced persistent threats may be answered by using big data analysis. These techniques could play a key role in helping detect threats at an early stage, especially when it uses a sophisticated pattern analysis, that works on different heterogeneous data sources. Given the numerous number of APT attacks that organizations face today, an APT security protective framework has been presented in [26]. The proposed framework integrates deep and 3D defense strategies. To protect against APT attacks, the system classifies data based on the level of confidential data. Botnet attacks is also another area where big data and machine learning techniques are deployed in. The work in [27] studied techniques for mitigating botnet attacks by using big data Analytics. The Advanced Cyber Defense Center (ACDC) orchestrated the sharing of gathered cybersecurity information on botnet attacks with the aim of defending through botnets. The work [28] proposed an architecture to address the current issue of botnet detection. They explored the possibility of employing a Self Organizing Map as an unsupervised learning approach to label unknown traffic. Financial sector is another area where big data analytics is used to prevent malicious actions or cyber attacks. The work in [29] studied using data fusion and visualization techniques in Network Forensic Analysis. Also, Cybersecurity Insurance (CI) is becoming more popular because of the increase of loss mitigation for cyber incidents for financial firms. Big data has now been employed in cybersecurity insurance, and the work in [30] proposed a framework which uses a big data approach in CI to analyze cyber incidents to gain insights in order to make better strategic decisions based on the information gathered. [31] investigated privacy and security issues associated with the sharing of financial data between institutions.

The work in [32] studied a novel Network Functions Virtualization-based (NFV) cybersecurity framework for providing security-as-a-Service in an evolved telco environment. The framework is known as SHIELD. This framework leverages BDA for detecting and mitigating threats in real time. [33] studied the idea of the construction of security monitoring systems for Internet of Things, which is based on parallel processing of data using the Hadoop platform. The proposed systems architecture has different components for the collection of data, storage of data, normalization and analysis, and visualization of data. Storage of data is done on Hadoop to improve the reliability and efficiency of processing of data requests. The work in [34] proposed a Security Information Management (SIM) enhancements using BDA. They devised a blueprint for a big data enhanced SIM, and field tested it using real network security logs. The work in [35] proposed a big data analytics model for protecting virtualized infrastructure in cloud computing. A Hadoop Distributed File system was used for the collection and storage of network logs and application logs from a guest virtual machine. Attack features were then extracted using graph-based event correlation and MapReduce parser identification of the potential paths of attack. A two-step machine learning algorithm using logistic regression and belief propagation were then applied to determine the presence of attacks. SIEM is an important tool in cybersecurity information analytics and a good source of

TABLE 1
Securing Big Data

Method/reference	Goals	Source of Data	Tools/Technologies
Security threats for big data [23]	Mitigating Phishing Attacks	Enron E-mail Dataset	Enronic Software
A big data architecture for security data [24]	Defend Against Spam and Phishing	pcap files, logs from honey net	Hadoop, Spark
Data mining methods for detection of malicious executables [55]	Detect malicious malware	Malicious and benign executable binaries	Machine Learning and Data Mining Algorithms
A practical solution to improve cybersecurity [53]	Security monitoring tool	Network dataset	Data Mining Techniques, High Functioning autistic graduates
Automate Cybersecurity Data Triage [54]	Help security analysts with data triage	The operation traces of security analysts on IDS logs and Firewalls	Data modeling and mining Techniques, Humans
Analyzing and Predicting Security Event Anomalies in BDA Deployment [36]	Improve SIEM by adding important features.	Traditional SIEM systems	Data Mining, Graph Analytics
Network Information Security on Big Data [26]	Advanced Persistent Threat Detection	Network data set	Big Data Analytics, Network event collection techniques, Big Data correlation analysis
Big Data machine learning and graph analytics [56]	Combining batch and stream data processes for efficiency reasons	Hetereogeneous Big Data (any type of data)	Lambda architecture
SIM in light of Big Data [34]	Cyber attack detection	Security logs	Machine learning techniques
Data fusion & visualization [29]	Network forensic investigation	Network logs	Data fusion techniques, Visualization, Self Organizing Map
Owlsight: Platform for real-time detection and visualization of cyber threats [20]	Real time detection and visualization of threats	Heterogeneous network data	Big Data Analytics, Web services, Data visualization
Predicting and fixing vulnerabilities before they occur: a Big Data approach [41]	Proactive Defense (Prevention better than cure approach)	Heterogeneous network data	Big Data Analytics techniques, Machine Learning
Machine learning classification model [45]	Network Intrusion Detection System in Android phones	Android data	Machine Learning Algorithms
A Big Data architecture for large scale monitoring [42]	Intrusion detection and prevention systems	NetFlow records, HTTP traffic and honeypot data	Shark, Spark, Machine Learning algorithms
A Scalable Meta-Model for Big Data Security Analysis [43]	Detect network anomaly at per flow level rather than the usual per packet level which tends to bring scalability issues	Network data	Machine learning and Data Mining Algorithms
Network security and anomaly detection [44]	Intrusion Detection System	Network flow Data	Spark, Cassandra, Machine Learning Algorithms
SHIELD: A novel NFV-based cybersecurity framework [32]	Security as a Service(SecaaS) to protect applications on Software	Heterogeneous cybersecurity data	Big Data Analytics, Machine Learning
Security evaluation of RC4 using Big Data analytics [37]	Analyzing the security of RC4	RC4 Algorithm	MapReduce, Big Data Analytics
Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing [35]	Big Data analytics model for protecting virtualized infrastructure in cloud computing	Network logs and application logs from a guest virtual machine	Machine Learning algorithms
Big Data security analysis approach using computational intelligence techniques [57]	Deduce the security status of the desktop and sources and causes of security breaches	Log file of Windows Firewall	Computational intelligence techniques
Data analytics on network traffic flows for botnet behaviour detection [28]	Issue of botnet detection	Network Traffic Data	Self Organizing Map as an unsupervised learning approach to label unknown traffic

TABLE 2
Research on Access Control and Encryption Techniques on Big Data

Method/reference	Problem	Solution
Computing on masked data: a high performance method for improving Big Data veracity [59]	Data not encrypted during Transmission	Improving on FHE by decreasing overhead
A faster fully homomorphic encryption scheme in Big Data [60]	Data not encrypted during Transmission	Improving on FHE by reducing public key size
A Data Masking Scheme for Sensitive Big Data Based on Format-Preserving Encryption [62]	Retain the original format of the plaintext instead of unreadable binary string during transmission	Format-Preserving Encryption (FPE) data masking scheme that chooses various FPE algorithms depending on the type of data and what needs to be done
Big Data Privacy Using Fully Homomorphic Non-Deterministic Encryption [61]	Data Security in the cloud during transmission	Fully Homomorphic non-deterministic encryption
Securing Personal Health Records in Cloud Computing: Patient-Centric and Fine-Grained Data Access Control in Multi-owner Settings [63]	Securing Personal Health Records in the cloud	Attribute-Based Encryption to ensure that each patient has a unique key based on his/her attributes
A Fine-Grained Access Control Scheme for Big Data Based on Classification Attributes [64]	Shortcomings of ABE encryption in cloud data storage services	Improve on the shortcomings of ABE by taking into account the relationships among the attributes
A digital envelope approach using attribute-based encryption for secure data exchange in IoT scenarios [65]	Improving Big Data Security	Better security by combing the flexibility of attribute-based cryptography and the efficiency of symmetric cryptography
CryptMDB: A practical encrypted MongoDB over Big Data [66]	Encryption of user data and in achieving privacy protection	Encrypted MongoDB which utilizes a homomorphic asymmetric cryptosystem
BigCrypt for Big Data encryption [67]	Overcome the limitation of asymmetric encryption techniques	BigCrypt(uses a probabilistic approach to Pretty Good Privacy Technique)
A Multi-level Intelligent Selective Encryption Control Model for Multimedia Big Data Security in Sensing System with Resource Constraint [74]	Security issues of heterogeneous, multimedia Big Data under resource constraints	Proposed a SAFE encryption scheme to replace old encryption models
Secure sensitive data sharing on a Big Data platform [68]	Securing the sharing of sensitive data on Big Data platform	Used proxy re-encryption algorithm based on heterogenous ciphertext transformation
Big Data protection via neural and quantum cryptography [69]	Protecting data	Using neural and quantum cryptography
Novel group key transfer protocol for Big Data security [70]	Secure group communication on Big Data,	Efficient group key transfer protocol using Diffie-Hellman key agreement
Double-Hashing Operation Mode for Encryption [71]	Cryptanalysis attacks	Used double hashing instead of a single hash
Enhancement CAST block algorithm to encrypt Big Data [72]	Enhancement of the cast block algorithm,	Use of one S-box instead of 6 to make it more dynamic
Less: Big Data sketching and encryption on low power platform [73]	Reducing and encrypting the processing of Big Data on low power platform	Light-weight Encryption using Scalable Sketching
Policy enforcement for Big Data security [76]	Privacy policy for Big Data security	Analyzes data, extracts the privacy policies, identifies sensitive data, then fragmentation algorithm executed on sensitive data
Managing the privacy and security of e-health data [75]	Privacy and security protection of clinical data	Art encryption scheme and attribute based authorization framework

data. The tool developed in [36] analyzes big data (gotten from SIEM) of a Fortune 500 company in order to gain insights about security threats through anomaly detection. They highlight the importance of graph analytics when it comes to intuitively understanding of business needs. Based on this, they apply graph analysis in anomaly detection by

TABLE 3
Research on Alternative Approaches to Securing Big Data

Method/reference	Problem	Solution
A framework for providing security to Personal Healthcare Records [81]	Securing personal health records	Framework that classifies data based on societal importance and sensitivity levels
A novel data security framework using E-MOD for Big Data [82]	Securing important attributes of Big Data	Ranking algorithm to determine attributes and data masking to protect them
A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With Pairing-Based Cryptography [85]	Securing Multimedia Big Data (MBD) in the healthcare cloud	Use fog computing to store Decoy Multimedia Big Data (DMBD)
A space-and-time efficient technique for Big Data security analytics [98]	Space and time issues of Big Data	Bloom filter and its variants
Another Look at Secure Big Data Processing: Formal Framework and a Potential Approach [90]	Protecting both the data and the program that processes the data	Hiding operations performed using steganography and FHE
Attribute relationship evaluation methodology for Big Data security [83]	Attribute selection method for protecting the value of Big Data	Determining attributes that have higher relevance using a ranking algorithm
Real time Big Data analytic: Security concern and challenges with Machine Learning algorithm [95]	real time Big Data analytics and its security challenge,	use of machine learning for protection of Big Data
Research on Network Security Visualization under Big Data Environment [96]	Visualizating network security under the Big Data environment	Use of radial traffic analyser and SRNET
Secure and private management of healthcare databases for data mining [99]	Secure and private management and mining of data in health care	Executing SQL queries on encrypted data and the return differentially-different answers on the outsourced databases
Secure Distribution of Big Data Based on BitTorrent [100]	Security issues accompanying P2P Big Data transmission avenues	Scheme for secure and efficient distribution of Big Data on BitTorrent networks using bittorrent protocols
Secure multimedia Big Data sharing in social networks using fingerprinting and encryption in the JPEG2000 compressed domain [87]	Protecting multimedia Big Data distribution in social networks	Homomorphic encrypted domain for fingerprinting by means of social media network analysis
Security in Big Data of medical records [86]	Securing Big Data in medical applications	Data hiding, image cryptography and stenography
Security-aware efficient mass distributed storage approach for cloud systems in Big Data [91]	Data storage in cloud computing could be accessed by cloud operators and therefore comprise information privacy and security	Splitting and separating the stored data
Security-as-a-service in Big Data of civil aviation [101]	Data protection and privacy preserving services architecture in civil aviation	Civil aviation security data authentication through OpenSSL identity and attribute-based authorization
Towards Early Detection of Novel Attack Patterns through the Lens of a Large-Scale Darknet [97]	Early identification of attacks using the darknet	Itemset mining engine to explore regularities in attack, then machine learning algorithms (clustering) to determine attack patterns and predict attacks
Big Data analysis based security situational awareness for smart grid [88]	Disadavantage of using traditional security framework for protection of the smart grid	Security awareness mechanism based on the analysis of Big Data in the smart grid
Big Data security hardening methodology using attributes relationship [84]	Protection of Big Data using a security hardening methodology	Makes use of attribute relationships to achieve it

TABLE 3
(*Continued*)

Method/reference	Problem	Solution
On the Future of Information: Reunification, Computability, Adaptation Cybersecurity, Semantics [89]	Problem of software vulnerability and the accumulation of unprocessed information in Big Data	Complete elimination of human intervention
Privacy preserving large scale DNA read-mapping in MapReduce framework using FPGAs [92]	Running BLAST algorithm in a secure manner in MapReduce framework using cloud computing	a Field programmable gate arrays (FPGA) based solution and a bitstream encryption mechanism
Efficient privacy-preserving dot-product computation for mobile Big Data [102]	Secure privacy-preserving scheme in mobile Big Data	Privacy-preserving dot product
Privacy-Preserved Multi-Party Data Merging with Secure Equality Evaluation [103]	Merging of encrypted data	Data anonymization technique that ensures privacy in the collection and merging of data and secures multiparty sharing of data without the involvement of third parties
Proposition of a method to aid Security Classification in Cybersecurity context [93]	Managing security classification	Assessing the risk behind various applications and providing an explanation of the ability of the application to protect data using a specific security classification level
Toward a cloud-based security intelligence with Big Data processing [104]	Cloud based security intelligence system for Big Data processing	Highly scalable plugin based solution that monitors Big Data systems in real time and therefore reducing the impact of attacks or threats on a distributed infrastructure
Research about New Media Security Technology Base on Big Data Era [105]	Security threats of the new Big Data in digital era	“Blocking as loose” technology for protection, intelligent cleaning of new media Big Data, and mining of Big Data in a safe manner.
Big Data Model of Security Sharing Based on Blockchain [106]	Model for security sharing based on blockchain technology to address trust issues often associated with circulation of data	blockchain and smart contract
Big Data for Cybersecurity: Vulnerability Disclosure Trends and Dependencies [107]	effective vulnerability management for organizations dealing with Big Data	proactive Big Data vulnerability management model based on rigorous statistical models with the capability of simulating anticipated volume and dependence of vulnerability disclosures
New approach for load rebalancer, scheduler in Big Data with security mechanism in cloud environment [108]	Rebalancing and scheduling of loads in Big Data environment,	Proposed scheme uses a load balancing algorithm that merges with MD5 and DES encryption algorithm
Hadoop eco system for Big Data security and privacy [94]	Secure and maintain the privacy of Big Data	Four encryption techniques. Using a buffer system where the buffer stores information whilst the system works on the previous data stored in order to prevent information loss

adding additional important capabilities of existing tools to their new tool, and then to visualize the network ins and outs. Finally, another use case of big data for security reasons involves a method for analyzing the security of RC4 [37]. Since attacks are diverse and come in multiple forms, BDA has been used as a cybersecurity tool to mitigate those attacks.

An area in cybersecurity where big data is used a lot is in Intrusion detection and prevention systems (IDS/IPS) research. Intrusion attempts are done to usually access information, interfere with the information or to tamper with a system thus making it unreliable and unusable. The IDS concept has been around for two decades but has recently seen a dramatic rise in the popularity and incorporation into the overall information security infrastructure [38]. IDSs are used to determine if there has been a breach or an interference in the network [39]. An IDS is often regarded as a second-line security solution after authentication, firewall, cryptography, and authorization techniques. Similarly, IPS can be classified into two categories: Network-based IPS and Host-based IPS. In network intrusion, prediction and detection is time sensitive, and needs highly efficient big data technologies to deal with problems on the fly [40]. This ensures a proactive rather than a reactive approach to cybersecurity. [41] approached this problem by developing a Proactive Cybersecurity (PCS) system. The PCS is a layered modular platform that makes use of big data collection and processing techniques to a wide variety of unstructured data to identify and thwart cybersecurity attacks. The PCS has a Targeted Vulnerability Predication (TVP) subsystem for detecting threats. Additionally, the model makes use of an Architectural Vulnerability Detection (AVD) subsystem and a risk analysis and recommender (RAR) subsystem for aiding identification and analysis of the identified risks (e.g., [16]). The work in [42] also proposed an architecture that handles IDS/IPS issues in a network. Their architecture stores and manages data from heterogeneous sources and also tries to find insights in the data. DNS data, NetFlow records, HTTP traffic and honeypot data were used in the research. Their approach however only provides offline analysis. Yang [43] presented an alternate approach that detects network anomaly at per-flow level rather than the usual per packet level which tends to bring scalability issues. They build a meta model for a number of machine learning and data mining algorithms. [44] also proposed a network security and anomaly detection framework for the big data systems for Network Traffic Monitoring and Analysis (NTMA) applications. Their framework is known as Big-DAMA. Big-DAMA is a very flexible Big Data Analytics framework (BDAF) that can perform analysis and storage of huge amounts of both heterogeneous structured and unstructured data. Big-DAMA also has batch and stream processing capabilities. Additionally, Big-DAMA utilizes Apache Spark Streaming for stream based analysis and for batch analysis, it uses Spark. For query and storage, it uses Apache Cassandra. Several machine learning algorithms are implemented by Big-DAMA for anomaly detection and network security. Big-DAMA was applied to various network attacks and anomaly detection. It was found to have the ability to speed computations by a factor of 10 in comparison to Apache Spark cluster. Security monitoring using big data has also been extended to other avenues. The work done in [45] also propose a Machine Learning model for Network-based Intrusion Detection Systems in order to detect the network security threats. Different types of ML classifiers are built using data-sets containing the labeled instances of network traffic. The focus of this research was to detect Android threats and give awareness and popularity to the users. This model can be integrated with traditional detection systems to detect advanced threats and reduce false positives. Thus, machine learning models are an essential part of BDA and have especially been used extensively in network anomaly detection.

2.2 Machine Learning (ML) in Cybersecurity

BDA and machine learning models go hand in hand. To provide security by deriving actionable insights, ML algorithms are needed to learn from the data. ML algorithms fall broadly into three categories: supervised learning, unsupervised learning and semi-supervised learning (which is a combination of supervised and unsupervised learning). The primary differentiator between supervised and unsupervised learning lies in the nature of the data that each uses. Unsupervised learning algorithms are used on data in which the outcome of each training sample is not known. A classic example is in malware detection. To achieve this, we extract the features from malware dataset and find groupings or similarities of the malware. The model uses the features of the data set to find its own groupings. Techniques that are used for unsupervised learning malware analysis are usually clustering algorithms and Principal Components Analysis (PCA). Supervised learning algorithms are trained on data in which the outcome of each training sample is already known. Some techniques used for doing supervised learning are linear and logistic regression, support vector machines, random forests and neural networks which are have commonly been re-branded as deep learning. Deep learning algorithms are very useful for analyzing large amounts of unsupervised data with high variety, which gives it potential in analyzing network data for intrusion detection, especially when it comes to NIDS [46], [47]. [48] tackled this issue when they used a deep learning technique called Self-taught Learning(STL) on the NSL-KDD dataset for intrusion detection on a network.

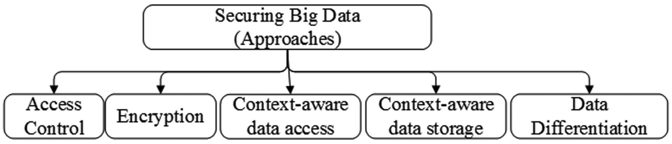
Securing Big Data (Approaches)
Access Control
Encryption
Context-aware data access
Context-aware data storage
Data Differentiation

Fig. 5. Securing the big data.

However, deep learning has some challenges in big data [49]. Its adaptability can be used as a vulnerability when attackers exploit the Machine Learning models. Adversarial examples [50] are machine learning inputs specifically designed to trick the ML model into producing a different output. Various works that have been done on this area try to refine the models [51]. [52] however propose a different approach to detecting adversarial examples. This approach is called feature squeezing which involves the reduction of the search space available to an adversary by merging samples that correspond to many different feature vectors in the original space into a single sample. With the advancement of Generative Adversarial Networks and big data, attackers are using artificial intelligence to circumvent some of the machine-learning automated processes. In lieu of this, a more effective approach is the merging of human and machine elements. Vimod [53] proposed an approach were humans and machine collaborate together. They used high-functioning autistic graduates with specific attributes to monitor networks and network flows. The other work [54] that incorporated the use of big data to assist humans studied data triage, and how helpful it is in identifying true attack patterns in a noisy data. This approach tries to automatically generate data triage automaton by tracing the actions of security analysts. This approach is different from existing data triage automaton like SIEM, because unlike SIEM, which requires analysts to manually generate event correlation rules, their approach mines data triage rules out of cybersecurity analysts' operation traces. It can be sheen that attackers are using artificial intelligence to trick ML models. Human and machine working together is one of the effective ways to combat these attacks.

3 Securing Big Data

Previous section presented how security can be achieved with big data. This section presents how to secure big data against different attacks. Typical techniques for securing big data are shown in Fig. 5. When data gets really big, securing it becomes really difficult. In [58], authors studied the security issues associated with big data and cloud computing. They identified the fact that most organizations outsource database in the form of big data in to the cloud. Cloud computing however still has many risks associated with it. The goal in [58] was to find security vulnerabilities in the cloud in order to inform vendors about recent vulnerabilities. They noted that confidentiality, integrity and availability in that order as the most important security issues a cloud provider faces. Confidentiality in this scenario would mean the protection of data against unauthorized interference or usage. Integrity would be the prevention of unauthorized and improper data modification. Availability would be akin to data recovery from hardware, software and system errors, and also from data access denials. However, confidentiality is the most important aspect when it comes to big data protection. Several data confidentiality techniques exist with the most notable ones being access control and encryption.

3.1 Access Control and Encryption Techniques for Big Data

Encryption and access control are similar in the sense that they are both synonymous with privacy and prevention. A notable difference however is that, encryption usually deals with the confidentiality of data. Data can be available to either a trusted or untrusted entity. Encryption ensures that only authorized trusted entities can view the data. Access control however tries to limit the access to data. The data limitations usually happens amongst trusted parties. For this reason, encryption techniques have to be stronger than access control techniques. Encryption imposes very strong limitations over data confidentiality. However, encryption is not an easy task. It tends to be computationally expensive and it has scalability issues (many users requiring access to the same data). Access control tends to be more flexible, and is easier to implement. When Big Data is transmitted to the cloud, a security issue emerges. Most organizations would not want their data in the hands of another organization, thus the need for encryption. A common approach is the use of data masking schemes. When the data is transmitted, it is not encrypted because the approaches used to transmit the data requires that the data be decrypted. This exposes the data to attacks. Confidentiality breach is the biggest threat to big data thus the encryption could be used as the primary big data protection technique.

In [59], authors studied the data transmission issues. They proposed computing on masked data to solve this. They proposed an incremental work to improve upon the already existing Fully Homomorphic Encryption (FHE) and other data masking techniques by decreasing the over head associated with other FHE techniques. The work in [60] also tried to improve on a Fully Homomorphic Encryption scheme for big data. It attempted to do this by reducing the public key size with the aim of making their scheme more efficient. The work in [61] proposed a model for the protection of data privacy using a fully homomorphic non-deterministic encryption. The proposed data protection model ensured the prior encryption of data before it was transmitted and therefore avoidance of the loss of data. The proposed system however only accepted numerical data. The output from the system was a result gotten from the computation of the encrypted data which is similar to that of the plaintext. In the future, the authors will look into the improvement of this anonymized data protection approach. The work in [62] explored the use of Format-Preserving Encryption (FPE) data masking scheme for voluminous data. The approach chooses various FPE algorithms depending on the type of data and what needs to be done. Spark framework was used. The authors chose FPE encryption technique because the ciphertext of FPE will still retain the original format of the plaintext instead of unreadable binary string. The ciphertext will now not contain any sensitive information. This approach however has its drawbacks. The encryption speed is slow when compared to other traditional symmetric algorithms. In one FPE method call, the algorithm calls the block cipher many times thereby making it inefficient. Another commonly used encryption is Attribute Based Encryption (ABE). The work in [63] presented a framework for fine-grained data access control to Personal Health Records (PHR) in the cloud that uses Attribute-Based Encryption as an encryption method to ensure that each patient has a unique key based on his/her attributes. The data could be accessed under multi owner settings. It was not only free of errors, but also protected the data from malicious parties aiming at deceiving the data users. In another paper involving ABE, Yang et al. [64] addressed some of the shortcomings of ABE encryption in cloud data storage services. They proposed a variant of ABE which is a novel distributed, scalable and fine-grained access control scheme based on the classification attributes of the cloud storage object. Their goal was to improve on the shortcomings of ABE by taking into account the relationships among the attributes. The work in [65] investigated a hybrid approach that combines symmetric cryptography and ABE to secure big data. They wanted to combine the flexibility of attribute-based cryptography and the efficiency of symmetric cryptography. They use Ciphertext-Policy Attribute-Based Encryption (CP-ABE) and AES encryption. In another form of big data encryption scheme, the work in [66] proposed an encrypted MongoDB which utilizes a homomorphic asymmetric cryptosystem which can be used for the encryption of user data and in achieving privacy protection. Thus, the FPE, FHE and ABE are the more popular researched big data encryption techniques.

A model for encrypting both symmetric and asymmetric data was presented in [67] which sought to overcome the limitation of asymmetric encryption techniques such as key exchange problem and the limited size of data and which in turn made it irrelevant for big data applications. Their proposed technique was known as BigCrypt which uses a probabilistic approach to Pretty Good Privacy Technique (PGP). BigCrypt encrypts the message with a symmetric key and encrypts the symmetric key using a public receiver key which is then attached to the message. The message is then sent. At the receiver end, the symmetric key is extracted and then asymmetrically decrypted and used for decrypting the main message. The proposed model was tested on local, web, and cloud server and was found to be efficient. Furthermore, a framework for securing the sharing of sensitive data on a big data platform was proposed in [68]. Sharing sensitive data securely reduces the cost of providing users with personalized services in addition to providing valueadded data services. The proposed scheme secures the distributed data, securely delivers it, stores it, and ensures secure usage. Semi-trusted big data is also destroyed. The proposed scheme uses a proxy re-encryption algorithm that is based on heterogeneous cipher-text transformation. The scheme also utilizes a user process protection method based on a virtual machine monitor that supports other system functions. This framework ensures data security while ensuring it is shared safely and securely. Sharma and Sharma in [69] discussed the protection of big data using neural and quantum cryptography. Neural cryptography incorporates the concept of artificial neural networks with classical cryptographic algorithms while quantum cryptography makes use of the phenomenon of quantum physics for securing communications. The authors also provided a comparative analysis between quantum and neural cryptography based on the methodologies that both techniques employ. From the analysis, the authors showed that a quantum computer makes use of quantum mechanisms for computation which are very powerful and can therefore crack complicated problems such as discrete logarithmic problem in a small duration. Neural key exchange protocol is also shown not to depend on any number theory. The analysis also indicates that neural networks probably have higher protection. The work in [70] proposed an efficient group key transfer protocol necessary for ensuring secure group communication on big data. The proposal does not use an online key generation centers (KGC) which is based on 3-LSSS (Linear secret sharing scheme) in that three modular multiplications are needed. Additionally, the protocol uses Diffie-Hellman key agreement. The proposed group key transfer scheme consists of two sections; two party secret establishment section and a section for the group session key transfer. The proposed group key transfer scheme was analyzed to verify its elements of key freshness, key confidentiality, and key authentication. Furthermore, the work in [71] proposed a new encryption scheme that can be used on big data that uses double hashing instead of a single hash. Double hashing they claim eliminates the threat of known cryptanalysis attacks. The work in [72] discussed primarily about the enhancement of CAST block algorithm for the security of big data. Their contribution to the enhancement of the cast block algorithm involved the use of one S-box instead of 6, and an approach to make it more dynamic. The work in [73] presented a framework that is Light-weight Encryption using Scalable Sketching(LESS) for reducing and encrypting the processing of big data on low power platform. This contains two kernels."sketching" and "sketch-reconstruction". Orthogonal Matching Pursuit (OMP) algorithm is implemented on the domain-specific Power Efficient Nano Cluster platform that acts as a hardware accelerator and ARM CPU for big data processing. Finally, the work [74], discuss the security issues of heterogeneous, multimedia big data. They tackle resource constraint issues such as limited computation and energy resources. They proposed data encryption models that deals with this issue by reducing the computation overload on weak nodes and by replacing the current encryption models with an improved version based on SAFE encryption scheme to improve it. The work in [75] mentioned a new approach for the privacy and security protection of clinical data through the use of the art encryption scheme and attribute based authorization framework.

For the access control and privacy of big data, the work in [77] presented a hybrid approach based framework that composes and enforces privacy policies to capture privacy requirements in an access control system. Gao et al. [78] presented a cloud security control mechanism based on big data. Cloud computing was observed to have increased the amount of data in the network. Due to this, big data leaks and losses occurred. Therefore, there was the need to provide the necessary level of protection. To that end, they conducted an analysis on big data, analyzed the current big data situation. Gupta et al. [79] proposed a security compliance model for big data systems. The model provides security and access control to big data systems at the initial stage. The proposed system has four models; the library, low critical log, high critical log, and a self-assurance system. The design of this system ensures real time analysis of big data. The initial level of security provided by the model is facilitated by its web directory and its self-assuring framework that identifies and differentiates genuine users and critical users. The relationship analysis tool of the users blocks users who are deemed not to be genuine. In [76], the authors proposed a framework for privacy policy for big data security. The proposed framework makes use of different techniques including security policy manager, fragmentation approach, encryption approach, and security manager. The characteristics of the privacy policy required flexibility, integration, customizability, and context-awareness. The framework works by receiving data from the customer and then analyzing it. It is then followed by the extraction of the privacy policy and finally the identification of sensitive data. Once sensitive data has been identified, a fragmentation algorithm was executed on the sensitive data. The security modules play the role of identifying sensitive data from non-sensitive data and then regulating its access. The work in [80] proposed a privacy protection technology and control mechanism for medical big data. The proposed framework has four main phases; the setup phase, Encrypt and Upload phase, Download phase, and Share File phase. The system first de-identifies the patient personal privacy data, encrypts it using digital signature mechanisms to protect data confidentiality and the authentication of the data. The communication security of the data in the system is protected using the Diffie-Hellman session key while the integrity of the medical records is protected using a digital signature scheme. Access control is not as big as it used to be due to the evolving threats landscape but is still an important research area in big data security today.

3.2 Alternative Approaches to Securing Big Data

Encryption and access control were the mainstream approaches for big data security. However, researchers have tried other approaches that may or may not involve some form of encryption. The nature of big data makes it difficult to protect everything. Some researchers have tried to determine the important parts of big data to protect those parts only. The work in [81] tried to tackle the issue of securing personal health records by proposing a framework that classifies data based on a person's societal importance and determining the sensitivity levels of the data. Furthermore, [82] tried to secure the attributes of big data that are really important/valuable because protecting everything is a difficult task. They use data masking to protect these high valued attributes. To determine the attributes that are of value, they use a ranking algorithm that prioritizes attributes for big data security. Authors in [83] proposed an attribute selection method for protecting the value of big data by determining attributes that have higher relevance using a ranking algorithm, and providing security measures. In the paper [84], the authors focused on the characteristics of big data and proposed the protection of big data using a security hardening methodology that makes use of attribute relationships. The relationship between the various attributes are expressed using nodes and edges. The proposed model works by limiting the attribute to protect value. The model works by first extracting all the attributes of the targeted big data. The nodes are then arranged circularly followed by the establishment of the relationship between the nodes. The relationship is then set based on either the domain specific criteria or the universal criteria. Finally, the protecting nodes are selected followed by the determination of how to protect the selected nodes. Thus, protecting everything in big data is hard. An easier approach is to find what is important and protect that part only.

Encryption has been used with other techniques as well. The work in [85] proposed a method to secure Multimedia Big Data (MBD) in the healthcare cloud by using a Decoy Multimedia Big Data (DMBD). The DMBD uses fog computing and a pairing based cryptography that will be used to secure the MBD. Fog Computing was utilized for the storage of the decoy files. In their method, the decoy files are retrieved at the onset unlike other methods that usually waits until there is an attack before the decoy files are called. Thus, both attacker and legitimate users both see a decoy file until the legitimacy is confirmed. Aynur in [86] presented a new technique for securing big data in medical applications. The methodology combines three major techniques that include data hiding, image cryptography and steganography. These techniques facilitate safe and denoised transmission of data. A stream cipher algorithm is used for encrypting the original image. Patient information is then embedded in the encrypted image by means of a lossless data embedding technique together with a key for hiding data to enhance the security of data. Steganography is then applied in embedded image with a private key. When the message gets to the receiver, it is decrypted using inverse methods in reverse order. Efficiently securing big data continues to become a difficult challenge because of big data's variety, volume and veracity issues. The ability to deal with space and time issues by correlating events would play an important role in securing big data. [87] discussed the growth of social media network such as Facebook and cloud computing, and how sharing of multimedia big data has become easier than ever. However, its increased used is faced with issues of piracy problems, illegal copying, and misappropriation. To address these challenges, the authors in this study proposed a system for protecting multimedia big data distribution in social networks. The scheme utilizes a Tree-Structured Harr (TSH) transform. In this scheme, a homomorphic encrypted domain for fingerprinting by means of social media network analysis is applied. The scheme aims at mapping hierarchical social networks into trees structure of the transform of TSH for coding, encrypting, and fingerprinting of JPEG2000. Finally, in [88], authors discussed the use of traditional security framework for protection of the smart grid comes with several disadvantages such as late detection of attacks when damage has already occurred. To address this problem, the authors in this study proposed a security awareness mechanism based on the analysis of big data in the smart grid. The model has three main parts which include the extraction of network security situation factors, network situational assessment and network situational prediction. The method works by integrating fuzzy cluster based analytical tools, reinforcement learning and game theory. The integration of these components facilitates security situational analysis in the smart grid. Simulation tests and experiments showed the proposed system to have high efficiency and low error rate.

Sometimes, we have to protect data from the people and the systems that interact with it. Pissanetzky [89] examined the problem of software vulnerability and the accumulation of unprocessed information in big data. According to the authors, these problems are created by human interventions. To solve these problems, the author proposed the complete elimination of human intervention. In this approach a causal set was taken as the universal language of all information and computations. Additionally, the author also proposed the confinement of the use of programming languages to the human interface and therefore a creation of an inner layer of mathematical code that is expressed as a causal set. Furthermore, this paper also includes experiments and computational verifications of the theory and proposed applications of this approach to science and technology, computer intelligence, and machine learning. Also, [90] researched on how to protect both the data and the program that processes the data while taking into consideration the big data processing requirements. They propose a model that aims to address the issue by hiding operations performed using steganography and FHE in order to meet the security requirements necessary to protect outsourced data. However, the user's computation cost is somewhat high and the solution does not apply to all applications. The work in [91] addressed the use of cloud computing and how it provides an organization with various services for meeting their various needs. However, data storage in cloud computing could be accessed by cloud operators and therefore compromise information privacy and security. In this respect, this study proposed an approach for splitting and separating the stored data on distributed cloud servers and therefore prevent access by cloud operators. The proposed model was known as Security-Aware Efficient Distributed Storage (SA-EDS) and was based on two algorithms; the Efficient Data Conflation (EDCon) algorithm and Secure Efficient Data Distributions (SED2) algorithm. These algorithms were tested and proved to be efficient. The authors of [92] proposed a Field Programmable Gate Arrays (FPGA) based solution for running BLAST algorithm in a secure manner in MapReduce framework using cloud computing. The proposed system protects data from cloud service provider (CSP) through leveraging on bitstream encryption mechanism and FPGAs tamper resistant property. The authors also put into consideration the risks that arise from keys distribution and propose countermeasures for handling it. The work in [93] studied an approach that assesses the risk behind various applications and provides an explanation of the ability of the application to protect data using a specific security classification level. The proposed method has three main components; Automatic Risk assessment of the Application, Automatic Generation of Criteria for storage of specific data, and Automatic Reporting. The report facilitates the recommendation of the appropriate security level. The work in [94] proposed a hadoop system that would both secure and maintain the privacy of big data. They tried to do this by using four encryption techniques randomly. However, these encryption techniques are time consuming, thus they proposed a buffer system where the buffer stores information whilst the system works on the previous data stored in order to prevent information loss.

Knowing the characteristics of the data is an important aspect of protecting the data. Singh [95] studied the value of real-time BDA and the security challenge that comes with protecting big data. Singh notes that, proper protection of big data should focus on volume, velocity, and variety of big data. Multilevel security for big data should be provided at the application, operating systems, and network levels. However, using the traditional protection mechanism is challenging for large volumes of data that is changing continuously. For this reason, Singh recommends the use of machine learning for protection of big data with focus on supervised and unsupervised learning. Yang [96] examined the visualization of network security under the big data environment. The authors first look at the 5V characteristics of big data including volume, velocity, variety, value, and visualize. These 5V features are then mapped onto network security data followed by a description of the visualization of the data security technology. The network visualization technology proposed include the use of radial traffic analyser and SRNET. They also proposed safety visualization using ClockMap and discussed diversified technologies for visualization of big data. With the increasing volume of big data, security and privacy issues also continue to increase. Peer to peer (P2P) protocols such as BitTorrent are now being used to widen the transfer of big data. However, this increase has also attracted widespread security challenges. Research indicate that P2P are sophisticated in data transfer but experience challenges when distributing big data. Ban et al. [97] presented a study on the early identification of attacks using the darknet. The system works by first exploring the regularities in communications from the attackers. This is achieved using an itemset mining engine. It then characterizes the activity level of each pattern of attack creating a time series. A clustering algorithm is then applied to extract the most prominent patterns of attack. The attack patterns are clustered into groups having similar activities. Visual hints on the relationship of the various attacks is then provided using a dimension reduction technique. Attacks that feature prominently are then the picked up for further analysis by experts. The authors showed that the proposed system was efficient in early attack detection.

The union of blockchain and big data will make sure that the data that is generated from the blockchain is trustworthy. This is because the provenance of the data is known. Also, the likelihood of the data being interfered with is very low. This is made possible through the blockchain's consensus mechanism and its secure cryptographic hash function which ensures data immutability. Data manipulation would require tremendous amount of hash power in order to be achieved. The centralized way of storing data is prone to data breaches and hacks [109]. This method is susceptible to single point of failure of problems as well. Distributed data storage tries to take data away from the hands of these centralized authorities, thereby taking away various security risks. The work in [106] proposed a model for security sharing based on blockchain technology to address trust issues often associated with circulation of data. The proposed model provides a credible platform for sharing data between data producers and demand parties though building a decentralized security system for the circulation of data. The security system is built using blockchain and smart contract. While blockchain technology ensures the traceability of data, the automated execution of smart contract provides the security for data security sharing. The decentralized architecture ensures the data provider does not suffer from the risks of sharing data from a centralized storage system. On the user's side, transparency in the collection of information is assured by the blockchain operation model and thereby bringing stronger user privacy protection. [110] proposed a system called MeDShare which is a blockchain based and provides data source auditing, and control for shared medical data in cloud repositories. The MeDShare system helps to transfer and share data from one source to another, and are recorded in a tamper-proof manner. The marriage of blockchain and Big Data is imminent as blockchain ensures data integrity.

The work in [102] proposed a secure privacy-preserving scheme using dot product in mobile big data. Privacy-preserving dot product has been used in data mining for a long time as it helps in curbing statistical analysis attacks. It is now being used in big data for its anonymous private profile matching. The paper was just an exploratory research on its use in mobile big data. There is however still room for further improvement. The work in [103] explored the idea of a data anonymization technique to support merging of encrypted data. The technique ensures the protection of privacy in the collection and merging of data and secures multi-party sharing of data without the involvement of third parties. The merging result as proposed in this study does not lead to the violation of the privacy of the individual. Additionally, the proposed mechanism allows for storage of different datasets from different parties in multiple third-party centers without leaking the identity of owners of that data. The anonymized data can be joined securely within a reasonable time. Experiments conducted by the authors indicated that 100,000 entries of data can be merged in about 1.4 seconds using the optimized secure merging procedure. To answer the question of how security classification can be managed on a system. In addition, the work in [104] proposed a cloud based security intelligence system for big data processing. The authors provide a highly scalable plug-in based solution that monitors big data systems in real time and therefore reduced the impact of attacks or threats on a distributed infrastructure. The solution proposed here was named Advanced Persistent Security Insights System (APSIS). APSIS works by taking advantage of a SIEM system including aggregation, correlation, alerting, and forensic analysis. This is exposed to big data but with security intelligence to provide accurate results. APSIS monitors all devices on the network that generate log files and therefore assures security. In the future, the authors aim at exploring the proof of concept to evaluate the robustness of the proposed architecture. The work in [105] started by looking at security threats of the new multimedia heterogeneous big data. The first threat was lack of effective mechanisms for the protection of this new media ownership as DRM is facing challenges. Second, there is lack of a clean environment for the consumption of new media. To overcome these challenges, Lu proposed the use of "blocking as loose" technology for protection, intelligent cleaning of new media big data, and the mining of big data in a safe manner. [98] summarized how bloom filter and its variants are used to secure big data. After various experiments, they concluded that, bloom filter can be used for efficiency reasons because there are space and time issues when it comes to analyzing and indexing big data which would in turn lead to better security analytics. The research work in [99] proposed a framework for secure and private management and mining of data that addresses both security and privacy issues in health-care data management especially in outsourced databases. The solution works by executing SQL queries on encrypted data and returning deferentially-different answers on the outsourced databases. Laplace mechanism are used to illustrate the computation of private queries. Private decision tree learning is also discussed. An experimental evaluation of the proposed solution shows the system incurs small communication and computation overhead. For this reason, the authors in this study [100] proposed a scheme for secure and efficient distribution of big data on BitTorrent networks. The proposed scheme is built inside the BitTorrent protocol and thus allowing the servers to regulate and trace user's behavior and sensitivity of data.

4 Existing Surveys on Big Data in Cybersecurity

Bertino [4] presented the security and privacy issues for big data concerning the confidentiality, privacy, and trustworthiness. In data confidentiality, the challenges identified were merging large number of access control policies and enforcing control policies in big data sources. Cybersecurity tasks such as user authentication, access control, and user monitoring are noted to be key in identifying threats and stopping them. The author noted that both security and privacy can be achieved by using advanced technologies such as cryptography. Mishra and Singh [5] examined security and privacy challenges associated with big data analytics for protecting database storage and transaction log files, and secure computations in distributed frameworks. The authors in [6] highlighted the benefits of big data analytics and reviewed security and privacy challenges in big data environments using various BDA tools such as Hadoop, MapReduce, and HDFS. Security and privacy challenges associated with big data environments were also listed as random distribution, security of big data computations, and access control. [7] examined big data emerging issues of security and privacy in relation to the use of big data analytic tools such as Hadoop. The work in [8] presented a review of big data security and privacy challenges while storing, searching and analyzing. In [9], the authors conducted a systematic literature review covering security and privacy for big data by categorizing approaches in terms of confidentiality, data integrity, privacy, data analysis, visualization, data format, and stream processing. Miloslavskaya et al. [10] examined the need for Security Operation Centres (SOCs) for organizations that want to achieve the highest protection for their data. The work in [111] looked at security intelligence centres (SICs) for processing of big data. The work in [112] proposed a framework which combined the techniques of security intelligence and big data analytics to support human analysts for prioritization. The work in [113] studied the security issues identified within the field of multimedia applications. In [11], Arora et al. performed a survey on big data and its security. The work in [114] highlighted the pros of big data, and then later tackles the challenges faced in China. In [115], Zou analyzed major issues associated with big data and especially the breach of personal information, the potential security risks, and the reduction of control rights of users over their personal information.

Mondek et al. [116] discussed security analytics in this era of big data and the reality of information security. Mahmood and Afzal [12] presented a review of big data analytics trends, tools, and techniques. The study of security analytics is motivated by the inadequacy of existing cybersecurity solutions to counteract cybersecurity attacks associated with big data. Jayasingh, Patra, and Mahesh [117] discussed security issues and challenges that faces security analysts in big data analytics and visualization. In [118], the authors discussed six changes in the information technology sector that they believe will be the game changers for the next 15 years. The work in [13], [14] presented security solutions for the big data in health-care industry. Health-care generates a lot of data from diverse sources and thus making it difficult to analyze. Similarly, in [119], Patil and Seshadri presented security and privacy issues in big data relating to the health-care security policies. The work in [120] summarized the current health-care security scenarios in big data environments in the USA.

The work in [15] put forward a model of big data security service for data providers, users, and cloud service providers. The work in [121] looked at opportunities, challenges, and security concerns associated with the use of big data in cloud computing. Furthermore, the work in [122] proposed integrated auditing for securing big data in the cloud. The authors presented their study by reviewing the characteristics of big data and security challenges in the cloud. The works in [123], [124], [125] proposed a security measure for big data, virtualization, and the cloud infrastructure and cloud based big data storage systems. Big data is making its way in the power industry. Smart grid has unique characteristics peculiar to it. The work in [126], [127] highlighted different articles that discuss the peculiarities of smart grid big data and how to properly handle it. Authors in [128] looked at security issues brought by big data applications in the telecommunication industry and especially associated with mobile network operators. In [129], authors surveyed three different techniques, namely homomorphic encryption, verifiable computation and multi-party computation. They discuss relevant security threats in the cloud, and a computation model that captures a large class of big data uses cases. The work in [130] studied the impact of security measures on the velocity of the big data system. This research found out that encryption is not an obstacle to the fast and efficient big data processing like it was before because of the introduction of new technologies. They recommended Encryption zones to be set as default in HDFS.

The work in [131] discussed the issues and challenges brought about by the big data deluge; data that is too big, too fast, and too diverse to the extent that are incompatible with the traditional database system. Paryasto et al. [132] presented the security challenges brought about by big data management through NIST risk management framework. The NIST SP800-30 framework provides a guide for conducting risk management on data. The work in [133] discussed the quality assurance for security applications of big data. The interest in quality assurance arises from the lack of confidence in the outcomes of big data applications. The risks in big data analytics arises out of lack of quality assurance. The work in [134] studied an on-line Cauchy based Clustering for cyber attack monitoring. In [135], authors classified big data during it's analysis phase in order to determine the security level of the data currently being analyzed. The work in [136] presented the various kinds of efforts that had taken towards for introducing a context-based information extraction using National Security Information Sources(NSIS) that enlist various kinds of knowledge inspired by natural activities of living things. The work in [137] showed the analysis of 79,012 articles that are published from the year 1916 to 2016 that relates to security and big data privacy.

In [138], the security of personal information on social media in this era of big data was presented. The study looked into the current situation of social network consumer privacy protection and attributed the security problem to personal information leakage and database defects. In [139], authors surveyed pre-processing techniques for data mining using conventional methods such as filtering, imputation, and embedding. The work in [140] discussed the challenges that exist in the era of wireless big data. Finally, the authors in [141] looked at ICT (considered to be the carrier of big data) supply chain security and big data. We provide a comprehensive study of recent research results by categorizing security with big data and big data security. In our work, we explore the role of big data in cybersecurity (as a tool and as an asset). We present up to date literature in this area and we highlight current and foreseeable challenges and trends in this field. We make it easier on readers by summarizing the problem each paper tried to solve and how they approached it in a tabular form.

5 Research Trends and Open Research Challenges

From the first ever virus known as the "creeper" and the first anti-virus made to neutralize it known as the "Reaper", the cybersecurity landscape has changed. The largest insider attack that ever occurred/happened for over a 30 year duration (1976 - 2006) and involved a former Boeing employee stealing intelligent info and handing them to China. Another well known insider threat was the Edward Snowden saga which involved the leakage of classified information from the NSA which resulted in the people distrusting the government. After, another major cyber security attack was Yahoo failing to report that the accounts of over 3 billion users have been jeopardized. Fast forward to 2017, the attack landscape is starting to shift again from data breaches to data being held for ransom. Ransomwares demanding payment (through cryptocurrency) condemn users to the erasure of their data if the ransom is not paid in beginning to gain traction (WannaCry and NotPretya Ransomware). The threat landscape is changing [142] and research trends need to change in other to combat these cybersecurity attacks. Typical trends, open research challenges and problems are shown in Fig. 6 and described below.

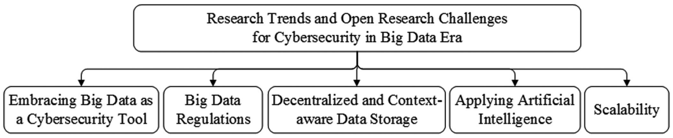
Research Trends and Open Research Challenges for Cybersecurity in Big Data Era
Embracing Big Data as a Cybersecurity Tool
Big Data Regulations
Decentralized and Context-aware Data Storage
Applying Artificial Intelligence
Scalability

Fig. 6. Typical trends, open research challenges and problems.

5.1 Embracing Big Data as a Cybersecurity Tool

Along with the data generated by IoT devices, the emergence of Bring Your Own Device (BYOD) has made organizations susceptible to various attack vectors. All these devices generate data. Thus organizations are starting to embrace BDA as a tool in their cybersecurity approach. Analyzing the data that passes through the network is essential to protecting the organization. However, some companies still have reservations on employing big data analytics as it tends to be an expensive undertaking. BDA also tends to be a complex field and requires expertise. Furthermore, employees are not comfortable with personal information gathered as this may involve tracking user activity. There are open challenges on how to differentiate the IoT system data, personal data and sensitive data and the protection of each of them using big data analytics.

5.2 Big Data Regulations

As a result of a myriad data breaches in recent times, new regulations such as Breach of Security Safeguards Regulations in Canada and Europe's General Data Protection Regulation have been implemented. A crucial aspect of the GDPR is the right to be forgotten, which gives an individual the power to enforce the deletion of any information pertaining to him/herself. A research trend we foresee here is self destructing data. Previous work has however been done on this. [143] propose an architecture that aims to solve the issue of personal data privacy. Their research was aimed at protecting the privacy of old data that has been stored on a centralized database which can then be re-used or re-surfaced. Their architecture made sure that copies of such data will become obsolete. This is a research area that might see a lot of growth in years to come, especially due to the emergence of blockchain and decentralized data storage. There are still challenges for big data regulation and policies including the situation where data leaves the organization for cloud storage.

5.3 Decentralized and Context-aware Data Storage

The most important commodity right now is data. The top companies GAFA (Google, Apple, Facebook, Amazon) have monopolized data, therefore bringing in the most revenue. New blockchain startups are now basing their business models on how to disrupt these monopolies by highlighting the value of data to the public. The selling point for these startups is that the data stored in a centralized fashion is susceptible to attacks (Facebook, Yahoo, and Equifax hackings) as evident in recent years. A distributed approach to storing data is the safer way to prevent attacks is what is being evangelized. This method is not susceptible to single point of failure problems as well. Companies such as the GAFAs store huge amount of data and they can correctly be termed as data silos. Distributed data storage try to take data away from the hands of these data silos, thereby taking away various security risks. Furthermore, the union of blockchain and big data will make sure that the data that is generated from the blockchain is trustworthy. There are ongoing research and open challenges on decentralized and context-aware data storage for big data.

5.4 Applying Artificial Intelligence

Artificial Intelligence based polymorphic malware is on the rise. Now, there is an application that can alter malware to trick machine learning antivirus software. In an experiment done by Endgame (a security company), they found out that AI has blind-spots that can be found out by other AI applications. This is evident as seen in Generative Adversarial Networks discovered by google researchers. This shows that organizations should not view machine learning as a fool proof way of defending against malware. More research work is needed in this area because of the rise of GANs. Also, an immediate approach to solve this would be to combine humans and AI in the malware detection approach. AI is not fool proof yet, and we see research trends gearing towards human in the Loop approaches to detect polymorphic malware.

5.5 Scalability for Cybersecurity Techniques in Big Data Era

In big data, protecting everything is hard. The easier approach is to find what is important and protect it. Traditional approaches for securing data might not work in a straightforward way. Thus, finding an optimal approach that is scalable for big data enabled systems is still an active research topic.

6 Conclusion

In this paper, we have surveyed state of the art literature on big data in cybersecurity. We segmented the work into two parts. The first part was research work involving the use of big data for security purposes. The second part is the research work done on securing big data. We present current trends on the use of BDA as security tool. We also addressed the role of machine learning in this area and some of the challenges machine learning has to overcome before it becomes an important feature in the cybersecurity toolkit. Furthermore, we discussed current literature on techniques used to secure big data. The confidentiality of big data is usually the main focus thus making encryption and access control techniques the main research areas when it comes to big data security. We also discussed the alternative approaches used to secure big data where the proposed approaches rely on other methods than encryption and access control in trying to secure other aspects of the CIA triad. We make it easier on readers by summarizing the problem each paper addresses and their approach to solve it in tabular form. Furthermore, we present future trends in big data security that we foresee, and the challenges associated with it.

Acknowledgments

This work is supported in part by the U.S. National Science Foundation (NSF) under grants CNS-1658972, CNS-1650831 and HRD 1828811. However, any opinion, finding, and conclusions or recommendations expressed in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the NSF.

References

[1] D. Laney, "3d data management: Controlling data volume, velocity and variety," *META Group Res. Note*, vol. 6, no. 70, p. 1, 2001.

[2] N. Miloslavskaya and A. Tolstoy, "Application of big data, fast data, and data lake concepts to information security issues," in *Proc. IEEE Int. Conf. Future Internet Things Cloud Workshops*, 2016, pp. 148–153.

[3] D. Rawat and K. Z. Ghafoor, *Smart Cities Cybersecurity and Privacy*. Amsterdam, The Netherlands: Elsevier, Dec. 2018.

[4] E. Bertino, "Big data-security and privacy," in *Proc. IEEE Int. Congr. Big Data*, 2015, pp. 757–761.

[5] A. D. Mishra and Y. B. Singh, "Big data analytics for security and privacy challenges," in *Proc. Int. Conf. Comput. Commun. Autom.*, 2016, pp. 50–53.

[6] Y. Gahi, M. Guennoun, and H. T. Mouftah, "Big data analytics: Security and privacy challenges," in *Proc. IEEE Symp. Comput. Commun.*, 2016, pp. 952–957.

[7] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, "Big data emerging issues: Hadoop security and privacy," in *Proc. 5th Int. Conf. Multimedia Comput. Syst.*, 2016, pp. 731–736.

[8] B. Matturdi, Z. Xianwei, L. Shuai, and L. Fuhong, "Big data security and privacy: A review," *China Commun.*, vol. 11, no. 14, pp. 135–145, 2014.

[9] B. Nelson and T. Olovsson, "Security and privacy for big data: A systematic literature review," in *Proc. IEEE Int. Conf. Big Data*, 2016, pp. 3693–3702.

[10] N. Miloslavskaya, A. Tolstoy, and S. Zapechnikov, "Taxonomy for unsecure big data processing in security operations centers," in *Proc. IEEE Int. Conf. Future Internet Things Cloud Workshops*, 2016, pp. 154–159.

[11] S. Arora, M. Kumar, P. Johri, and S. Das, "Big heterogeneous data and its security: A survey," in *Proc. Int. Conf. Comput., Commun. Autom.*, 2016, pp. 37–40.

[12] T. Mahmood and U. Afzal, "Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools," in *Proc. 2nd Nat. Conf. Inf. Assurance*, 2013, pp. 129–134.

[13] S. Rao, S. Suma, and M. Sunitha, "Security solutions for big data analytics in healthcare," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Eng.*, 2015, pp. 510–514.

[14] I. Olaronke and O. Oluwaseun, "Big data in healthcare: Prospects, challenges and resolutions," in *Proc. Future Technol. Conf.*, 2016, pp. 1152–1157.

[15] H.-T. Cui, "Research on the model of big data serve security in cloud environment," in *Proc. IEEE Int. Conf. Comput. Commun. Internet*, 2016, pp. 514–517.

[16] E. Damiani, "Toward big data risk analysis," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 1905–1909.

[17] C. Sinclair, L. Pierce, and S. Matzner, "An application of machine learning to network intrusion detection," in *Proc. 15th Annu. Comput. Secur. Appl. Conf.*, 1999, pp. 371–377.

[18] E. Chickowski, "A case study in security big data analysis," *Dark Reading*, vol. 9, 2012. https://www.darkreading.com/analytics/security-monitoring/a-case-study-in-security-big-data-analysis/d/d-id/1137299

[19] M. C. Raja and M. A. Rabbani, "Big data analytics security issues in data driven information system," *IJIRCCE*, vol. 2, no. 10, pp. 6132–6134, 2014.

[20] V. S. Carvalho, M. J. Polidoro, and J. P. Magalhães, "Owlsight: Platform for real-time detection and visualization of cyber threats," in *Proc. IEEE 2nd Int. Conf. Big Data Security on Cloud/IEEE Int. Conf. High Perform. Smart Comput/IEEE Int. Conf. Intell. Data Secur.*, 2016, pp. 61–66.

[21] Y. Yao, L. Zhang, J. Yi, Y. Peng, W. Hu, and L. Shi, "A framework for big data security analysis and the semantic technology," in *Proc. 6th Int. Conf. IT Convergence Secur.*, 2016, pp. 1–4.

[22] ProofPoint.com, "The human factor report people-centered threats define the landscape," 2018. https://cdn2.hubspot.net/hubfs/508286/blog-files/The%20Human%20Factore%20Report%202018.pdf

[23] T. Zaki, M. S. Uddin, M. M. Hasan, and M. N. Islam, "Security threats for big data: A study on enron e-mail dataset," in *Proc. Int. Conf. Res. Innovation Inf. Syst.*, 2017, pp. 1–6.

[24] P. H. Las-Casas, V. S. Dias, W. Meira, and D. Guedes, "A big data architecture for security data and its application to phishing characterization," in *Proc. IEEE 2nd Int. Conf. Big Data Security on Cloud/IEEE Int. Conf. High Perform. Smart Comput/IEEE Int. Conf. Intell. Data Secur.*, 2016, pp. 36–41.

[25] A. A. Cárdenas, P. K. Manadhata, and S. Rajan, "Big data analytics for security intelligence," pp. 1–22, September, 2013, Technical Report by Big Data Working Group of Cloud Security Alliance, https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf.

[26] W. Jia, "Study on network information security based on big data," in *Proc. 9th Int. Conf. Meas. Technol. Mechatronics Autom.*, 2017, pp. 408–409.

[27] B. G.-N. Crespo and A. Garwood, "Fighting botnets with cyber-security analytics: Dealing with heterogeneous cyber-security information in new generation siems," in *Proc. 9th Int. Conf. Availability Rel. Secur*, 2014, pp. 192–198.

[28] D. C. Le, A. N. Zincir-Heywood, and M. I. Heywood, "Data analytics on network traffic flows for botnet behaviour detection," in *Proc. IEEE Symp. Series Comput. Intell.*, 2016, pp. 1–7.

[29] H. Fatima, S. Satpathy, S. Mahapatra, G. Dash, and S. K. Pradhan, "Data fusion & visualization application for network forensic investigation-a case study," in *Proc. 2nd Int. Conf. Anti-Cyber Crimes*, 2017, pp. 252–256.

[30] K. Gai, M. Qiu, and S. A. Elnagdy, "A novel secure big data cyber incident analytics framework for cloud-based cybersecurity insurance," in *Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud/IEEE Int. Conf. High Perform. Smart Comput/IEEE Int. Conf. Intell. Data Secur*, 2016, pp. 171–176.

[31] K. Gai, M. Qiu, and S. A. Elnagdy, "Security-aware information classifications using supervised learning for cloud-based cyber risk management in financial big data," in *Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud/IEEE Int. Conf. High Perform. Smart Comput/IEEE Int. Conf. Intell. Data Secur*, 2016, pp. 197–202.

[32] G. Gardikis, K. Tzoulas, K. Tripolitis, A. Bartzas, S. Costicoglou, A. Lioy, B. Gaston, C. Fernandez, C. Davila, A. Litke, et al., "Shield: A novel nfv-based cybersecurity framework," in *Proc. IEEE Conf. Netw. Softwarization*, 2017, pp. 1–6.

[33] I. Saenko, I. Kotenko, and A. Kushnerevich, "Parallel processing of big heterogeneous data for security monitoring of iot networks," in *Proc. 25th Euromicro Int. Conf. Parallel Distrib. Netw.-Based Process.*, 2017, pp. 329–336.

[34] F. Gottwalt and A. P. Karduck, "Sim in light of big data," in *Proc. 11th Int. Conf. Innovations Inf. Technol*, 2015, pp. 326–331.

[35] T. Y. Win, H. Tianfield, and Q. Mair, "Big data based security analytics for protecting virtualized infrastructures in cloud computing," *IEEE Trans. Big Data*, 2017, vol. 4, no. 1, pp. 11–25, Mar. 2018.

[36] C. Puri and C. Dukatz, "Analyzing and predicting security event anomalies: Lessons learned from a large enterprise big data streaming analytics deployment," in *Proc. 26th Int. Workshop Database Expert Syst. Appl.*, 2015, pp. 152–158.

[37] C. Liu, Y. Cai, and T. Wang, "Security evaluation of rc4 using big data analytics," in *Proc. 7th IEEE Int. Conf. Softw. Eng. Serv. Sci.*, 2016, pp. 316–320.

[38] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (idps)," *NIST Special Publication*, vol. 800, no. 2007, 2007, Art. no. 94.

[39] S. Mukkamala, A. Sung, and A. Abraham, "Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools," *Enhancing Comput. Secur. Smart Technol.*, Rao V. (Ed.), CRC Press, USA, ISBN 0849330459, pp. 125–161, 2005.

[40] S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 4, pp. 70–73, 2014.

[41] H.-M. Chen, R. Kazman, I. Monarch, and P. Wang, "Predicting and fixing vulnerabilities before they occur: a big data approach," in *Proc. 2nd ACM Int. Workshop BIG Data Softw. Eng.*, 2016, pp. 72–75.

[42] S. Marchal, X. Jiang, R. State, and T. Engel, "A big data architecture for large scale security monitoring," in *Proc. IEEE Int. Congr. Big Data*, 2014, pp. 56–63.

[43] B. Yang and T. Zhang, "A scalable meta-model for big data security analyses," in *Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud/IEEE Int. Conf. High Perform. Smart Comput/IEEE Int. Conf. Intell. Data Secur.*, 2016, pp. 55–60.

[44] P. Casas, F. Soro, J. Vanerio, G. Settanni, and A. D'Alconzo, "Network security and anomaly detection with big-dama, a big data analytics framework," in *Proc. IEEE 6th Int. Conf. Cloud Netw.*, 2017, pp. 1–7.

[45] S. Kumar, A. Viinikainen, and T. Hamalainen, "Machine learning classification model for network based intrusion detection system," in *Proc. 11th Int. Conf. Internet Technol. Secured Trans.*, 2016, pp. 242–249.

[46] U. Fiore, F. Palmieri, A. Castiglione, and A. De Santis, "Network anomaly detection with the restricted boltzmann machine," *Elsevier Neurocomputing*, vol. 122, pp. 13–23, 2013.

[47] M. A. Salama, H. F. Eid, R. A. Ramadan, A. Darwish, and A. E. Hassanien, "Hybrid intelligent intrusion detection scheme," in *Soft Computing in Industrial Applications*. New York, NY, USA: Springer, 2011, pp. 293–303.

[48] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proc. 9th EAI Int. Conf. Bio-Inspired Inf. Commun. Technol.*, 2016, pp. 21–26.

[49] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, 2015, Art. no. 1.

[50] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[51] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint arXiv:1611.03814*, 2016.

[52] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017. https://machine-learning-and-security.github.io/papers/mlsec17_paper_52.pdf

[53] V. Patel, "A practical solution to improve cyber security on a global scale," in *Proc. 3rd Worldwide Cybersecurity Summit*, 2012, pp. 1–5.

[54] C. Zhong, J. Yen, P. Liu, and R. F. Erbacher, "Automate cybersecurity data triage by leveraging human analysts' cognitive process," in *Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud/IEEE Int. Conf. High Perform. Smart Comput/IEEE Int. Conf. Intell. Data Secur.*, 2016, pp. 357–363.

[55] M. G. Schultz, E. Eskin, F. Zadok, and S. J. Stolfo, "Data mining methods for detection of new malicious executables," in *Proc. IEEE Symp. Secur. Privacy*, 2001, pp. 38–49.

[56] H. H. Huang and H. Liu, "Big data machine learning and graph analytics: Current state and future challenges," in *Proc. IEEE Int. Conf. Big Data*, 2014, pp. 16–17.

[57] N. Naik, P. Jenkins, N. Savage, and V. Katos, "Big data security analysis approach using computational intelligence techniques in r for desktop users," in *Proc. IEEE Symp. Series Comput. Intell.*, 2016, pp. 1–8.

[58] K. Kaur, A. Syed, A. Mohammad, and M. N. Halgamuge, "An evaluation of major threats in cloud computing associated with big data," in *Proc. IEEE 2nd Int. Conf. Big Data Anal.*, 2017, pp. 368–372.

[59] J. Kepner, V. Gadepally, P. Michaleas, N. Schear, M. Varia, A. Yerukhimovich, and R. K. Cunningham, "Computing on masked data: a high performance method for improving big data veracity," in *Proc. IEEE High Perform. Extreme Comput. Conf.*, 2014, pp. 1–6.

[60] D. Wang, B. Guo, Y. Shen, S.-J. Cheng, and Y.-H. Lin, "A faster fully homomorphic encryption scheme in big data," in *Proc. IEEE 2nd Int. Conf. Big Data Anal.*, 2017, pp. 345–349.

[61] T. B. Patil, G. K. Patnaik, and A. T. Bhole, "Big data privacy using fully homomorphic non-deterministic encryption," in *Proc. IEEE 7th Int. Adv. Comput. Conf.*, 2017, pp. 138–143.

[62] B. Cui, B. Zhang, and K. Wang, "A data masking scheme for sensitive big data based on format-preserving encryption," in *Proc. IEEE Int. Conf. Comput. Sci. Eng./IEEE Int. Conf. Embedded Ubiquitous Comput.*, 2017, vol. 1, pp. 518–524.

[63] M. Li, S. Yu, K. Ren, and W. Lou, "Securing personal health records in cloud computing: Patient-centric and fine-grained data access control in multi-owner settings," in *Proc. Int. Conf. Secur. Privacy Commun. Syst.*, 2010, vol. 10, pp. 89–106.

[64] T. Yang, P. Shen, X. Tian, and C. Chen, "A fine-grained access control scheme for big data based on classification attributes," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. Workshops*, 2017, pp. 238–245.

[65] S. Pérez, J. L. Hernández-Ramos, D. Pedone, D. Rotondi, L. Straniero, and A. F. Skarmeta, "A digital envelope approach using attribute-based encryption for secure data exchange in iot scenarios," in *Proc. Global Internet Things Summit*, 2017, pp. 1–6.

[66] G. Xu, Y. Ren, H. Li, D. Liu, Y. Dai, and K. Yang, "Cryptmdb: A practical encrypted mongodb over big data," in *Proc. IEEE Int. Conf. Commun.*, 2017, pp. 1–6.

[67] A. Al Mamun, K. Salah, S. Al-maadeed, and T. R. Sheltami, "Bigcrypt for big data encryption," in *Proc. 4th Int. Conf. Softw. Defined Syst.*, 2017, pp. 93–99.

[68] X. Dong, R. Li, H. He, W. Zhou, Z. Xue, and H. Wu, "Secure sensitive data sharing on a big data platform," *Tsinghua Sci. Technol.*, vol. 20, no. 1, pp. 72–80, 2015.

[69] A. Sharma and D. Sharma, "Big data protection via neural and quantum cryptography," in *Proc. 3rd Int. Conf. Comput. Sustainable Global Develop.*, 2016, pp. 3701–3704.

[70] C. Zhao and J. Liu, "Novel group key transfer protocol for big data security," in *Proc. IEEE Adv. Inf. Technol. Electron. Autom. Control Conf.*, 2015, pp. 161–165.

[71] S. Almuhammadi and A. Amro, "Double-hashing operation mode for encryption," in *Proc. IEEE 7th Annu. Comput. Commun. Workshop Conf.*, 2017, pp. 1–7.

[72] F. A. Kadhim, G. H. Abdul-Majeed, and R. S. Ali, "Enhancement cast block algorithm to encrypt big data," in *Proc. Annu. Conf. New Trends Inf. Commun. Technol. Appl.*, 2017, pp. 80–85.

[73] A. Kulkarni, C. Shea, H. Homayoun, and T. Mohsenin, "Less: Big data sketching and encryption on low power platform," in *Proc. Conf. Des. Autom. Test Eur.*, pp. 1635–1638, 2017.

[74] C. Xiao, L. Wang, Z. Jie, and T. Chen, "A multi-level intelligent selective encryption control model for multimedia big data security in sensing system with resource constraints," in *Proc. IEEE 3rd Int. Conf. Cyber Secur. Cloud Comput.*, 2016, pp. 148–153.

[75] A. Soceanu, M. Vasylenko, A. Egner, and T. Muntean, "Managing the privacy and security of ehealth data," in *Proc. 20th Int. Conf. Control Syst. Comput. Sci.*, 2015, pp. 439–446.

[76] A. Al-Shomrani, F. Fathy, and K. Jambi, "Policy enforcement for big data security," in *Proc. 2nd Int. Conf. Anti-Cyber Crimes*, 2017, pp. 70–74.

[77] A. Samuel, M. I. Sarfraz, H. Haseeb, S. Basalamah, and A. Ghafoor, "A framework for composition and enforcement of privacy-aware and context-driven authorization mechanism for multimedia big data," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1484–1494, Sep. 2015.

[78] F. Gao, "Research on cloud security control mechanism based on big data," in *Proc. Int. Conf. Smart Grid Electr. Autom.*, 2017, pp. 366–370.

[79] A. Gupta, A. Verma, P. Kalra, and L. Kumar, "Big data: A security compliance model," in *Proc. Conf. IT Business Ind. Government*, 2014, pp. 1–5.

[80] N.-Y. Lee and B.-H. Wu, "Privacy protection technology and access control mechanism for medical big data," in *Proc. 6th IIAI Int. Congr. Adv. Appl. Informat.*, 2017, pp. 424–429.

[81] M. R. Islam, M. Habiba, and M. I. I. Kashem, "A framework for providing security to personal healthcare records," in *Proc. Int. Conf. Netw. Syst. Secur.*, 2017, pp. 168–173.

[82] R. Achana, R. S. Hegadi, and T. Manjunath, "A novel data security framework using e-mod for big data," in *Proc. IEEE Int. WIE Conf. Electr. Comput. Eng.*, 2015, pp. 546–551.

[83] S.-H. Kim, N.-U. Kim, and T.-M. Chung, "Attribute relationship evaluation methodology for big data security," in *Proc. Int. Conf. IT Convergence Secur.*, 2013, pp. 1–4.

[84] S.-H. Kim, J.-H. Eom, and T.-M. Chung, "Big data security hardening methodology using attributes relationship," in *Proc. Int. Conf. Inf. Sci. Appl.*, 2013, pp. 1–2.

[85] H. A. Al Hamid, S. M. M. Rahman, M. S. Hossain, A. Almogren, and A. Alamri, "A security model for preserving the privacy of medical big data in a healthcare cloud using a fog computing facility with pairing-based cryptography," *IEEE Access*, vol. 5, pp. 22313–22328, 2017.

[86] A. Unal, "Security in big data of medical records," in *Proc. Conf. IT Bus. Ind. Government*, 2014, pp. 1–2.

[87] C. Ye, Z. Xiong, Y. Ding, J. Li, G. Wang, X. Zhang, and K. Zhang, "Secure multimedia big data sharing in social networks using fingerprinting and encryption in the jpeg2000 compressed domain," in *Proc. IEEE 13th Int. Conf. Trust Secur. Privacy Comput. Commun.*, 2014, pp. 616–621.

[88] J. Wu, K. Ota, M. Dong, J. Li, and H. Wang, "Big data analysis based security situational awareness for smart grid," *IEEE Trans. Big Data*, vol. 4, no. 3, pp. 408–417, Sep. 2016.

[89] S. Pissanetzky, "On the future of information: Reunification, computability, adaptation, cybersecurity, semantics," *IEEE Access*, vol. 4, pp. 1117–1140, 2016.

[90] L. Xu, P. D. Khoa, S. H. Kim, W. W. Ro, and W. Shi, "Another look at secure big data processing: Formal framework and a potential approach," in *Proc. IEEE 8th Int. Conf. Cloud Comput.*, 2015, pp. 548–555.

[91] K. Gai, M. Qiu, and H. Zhao, "Security-aware efficient mass distributed storage approach for cloud systems in big data," in *Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud/IEEE Int. Conf. High Perform. Smart Comput/IEEE Int. Conf. Intell. Data Secur.*, 2016, pp. 140–145.

[92] L. Xu, H. Kim, X. Wang, W. Shi, and T. Suh, "Privacy preserving large scale dna read-mapping in MapReduce framework using fpgas," in *Proc. 24th Int. Conf. Field Programmable Logic Appl.*, 2014, pp. 1–4.

[93] G. Collard, E. Disson, G. Talens, and S. Ducroquet, "Proposition of a method to aid security classification in cybersecurity context," in *Proc. 14th Annu. Conf. Privacy Secur. Trust*, 2016, pp. 88–95.

[94] P. Adluru, S. S. Datla, and X. Zhang, "Hadoop eco system for big data security and privacy," in *Proc. IEEE Long Island Syst. Appl. Technol. Conf.*, 2015, pp. 1–6.

[95] J. Singh, "Real time big data analytic: Security concern and challenges with machine learning algorithm," in *Proc. Conf. IT Bus. Ind. Government*, 2014, pp. 1–4.

[96] T. Yang and S. Jia, "Research on network security visualization under big data environment," in *Proc. Int. Comput. Symp.*, 2016, pp. 660–662.

[97] T. Ban, S. Pang, M. Eto, D. Inoue, K. Nakao, and R. Huang, "Towards early detection of novel attack patterns through the lens of a large-scale darknet," in *Proc. Int. IEEE Conf. Ubiquitous Intell. Comput./Adv. Trusted Comput./Scalable Comput. Commun./Cloud Big Data Comput./Internet People/Smart World Congr.*, 2016, pp. 341–349.

[98] S. A. Alsuhibany, "A space-and-time efficient technique for big data security analytics," in *Proc. Saudi Int. Conf. Inf. Technol. (Big Data Analysis)*, 2016, pp. 1–6.

[99] N. Mohammed, S. Barouti, D. Alhadidi, and R. Chen, "Secure and private management of healthcare databases for data mining," in *Proc. IEEE 28th Int. Symp. Comput.-Based Med. Syst.*, 2015, pp. 191–196.

[100] L. Xiao, C. Xu, J. Qin, G. Qin, M. Zhu, L. Ruan, Z. Wang, M. Li, and D. Tan, "Secure distribution of big data based on bittorrent," in *Proc. IEEE 11th Int. Conf. Dependable Autonomic Secure Comput.*, 2013, pp. 82–90.

[101] W. Zhijun and W. Caiyun, "Security-as-a-service in big data of civil aviation," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2015, pp. 240–244.

[102] C. Hu and Y. Huo, "Efficient privacy-preserving dot-product computation for mobile big data," *IET Commun.*, vol. 11, no. 5, pp. 704–712, 2016.

[103] S. Q. Ren, T. H. Meng, N. Yibin, and K. M. M. Aung, "Privacy-preserved multi-party data merging with secure equality evaluation," in *Proc. Int. Conf. Cloud Comput. Res. Innovations*, 2016, pp. 34–41.

[104] K. Benzidane, H. El Alloussi, O. El Warrak, L. Fetjah, S. J. Andaloussi, and A. Sekkaki, "Toward a cloud-based security intelligence with big data processing," in *Proc. IEEE/IFIP Netw. Operations Manage. Symp.*, 2016, pp. 1089–1092.

[105] Z.-W. Lu, "Research about new media security technology base on big data era," in *Proc. IEEE 14th Int. Dependable Autonomic Secure Comput./14th Int Conf. Pervasive Intell. Comput/2nd Int. Conf Big Data Intell. Comput./Cyber Sci. Technol. Congr.*, 2016, pp. 933–936.

[106] L. Yue, H. Junqin, Q. Shengzhi, and W. Ruijin, "Big data model of security sharing based on blockchain," in *Proc. 3rd Int. Conf. Big Data Comput. Commun.*, 2017, pp. 117–121.

[107] M. Tang, M. Alazab, and Y. Luo, "Big data for cybersecurity: Vulnerability disclosure trends and dependencies," *IEEE Trans. Big Data*, p. 1, 2018. https://doi.org/10.1109/TBDATA.2017.2723570

[108] P. A. Dhande and A. Kadam, "New approach for load rebalancer, scheduler in big data with security mechanism in cloud environment," in *Proc. IEEE Int Conf Adv Electron. Commun. Comput. Technol.*, 2016, pp. 247–250.

[109] D. Puthal, N. Malik, S. P. Mohanty, E. Kougianos, and C. Yang, "The blockchain as a decentralized security framework," *IEEE Consum. Electron. Mag.*, vol. 7, no. 2, pp. 18–21, Mar. 2018.

[110] Q. Xia, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du, and M. Guizani, "Medshare: Trust-less medical data sharing among cloud service providers via blockchain," *IEEE Access*, vol. 5, pp. 14757–14767, 2017.

[111] N. Miloslavskaya, "Security intelligence centers for big data processing," in *Proc. 5th Int. Conf. Future Internet Things Cloud Workshops*, 2017, pp. 7–13.

[112] M. Marchetti, F. Pierazzi, A. Guido, and M. Colajanni, "Countering advanced persistent threats through security intelligence and big data analytics," in *Proc. 8th Int. Conf. Cyber Conflict*, 2016, pp. 243–261.

[113] Q. Jin, Y. Xiang, G. Sun, Y. Liu, and C.-C. Chang, "Cybersecurity for cyber-enabled multimedia applications," *IEEE MultiMedia*, vol. 24, no. 4, pp. 10–13, Oct.-Dec. 2017.

[114] Y. Mengke, Z. Xiaoguang, Z. Jianqiu, and X. Jianjian, "Challenges and solutions of information security issues in the age of big data," *China Commun.*, vol. 13, no. 3, pp. 193–202, 2016.

[115] H. Zou, "Protection of personal information security in the age of big data," in *Proc. 12th Int. Conf. Comput. Intell. Secur.*, 2016, pp. 586–589.

[116] D. Mondek, R. B. Blažek, and T. Zahradnickỳ, "Security analytics in the big data era," in *Proc. IEEE Int. Conf. Softw. Quality Rel. Secur. Companion*, 2017, pp. 605–606.

[117] B. B. Jayasingh, M. Patra, and D. B. Mahesh, "Security issues and challenges of big data analytics and visualization," in *Proc. 2nd Int. Conf. Contemporary Comput. Informat.*, 2016, pp. 204–208.

[118] A. Kott, A. Swami, and P. McDaniel, "Security outlook: six cyber game changers for the next 15 years," *Comput.*, vol. 47, no. 12, pp. 104–106, 2014.

[119] H. K. Patil and R. Seshadri, "Big data security and privacy issues in healthcare," in *Proc. IEEE Int. Congr. Big Data*, 2014, pp. 762–765.

[120] S. Chandra, S. Ray, and R. Goswami, "Big data security in healthcare: Survey on frameworks and algorithms," in *Proc. IEEE 7th Int. Adv. Comput. Conf.*, 2017, pp. 89–94.

[121] S. Anandaraj and M. Kemal, "Research opportunities and challenges of security concerns associated with big data in cloud computing," in *Proc. Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud)*, 2017, pp. 746–751.

[122] Y. Wang, B. Rawal, and Q. Duan, "Securing big data in the cloud with integrated auditing," in *Proc. IEEE Int. Conf. Smart Cloud*, 2017, pp. 126–131.

[123] S. Bahulikar, "Security measures for the big data, virtualization and the cloud infrastructure," in *Proc. 1st India Int. Conf. Inf. Process.*, 2016, pp. 1–4.

[124] A. Sharif, S. Cooney, S. Gong, and D. Vitek, "Current security threats and prevention measures relating to cloud services, hadoop concurrent processing, and big data," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 1865–1870.

[125] Z. Tan, U. T. Nagar, X. He, P. Nanda, R. P. Liu, S. Wang, and J. Hu, "Enhancing big data security with collaborative intrusion detection," *IEEE Cloud Comput.*, vol. 1, no. 3, pp. 27–33, Sep. 2014.

[126] J. Zhao, Y. Wang, and Y. Xia, "Analysis of information security of electric power big data and its countermeasures," in *Proc. 12th Int. Conf. Comput. Intell. Secur.*, 2016, pp. 243–248.

[127] J. Hu and A. V. Vasilakos, "Energy big data analytics and security: Challenges and opportunities," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2423–2436, Sep. 2016.

[128] C. Dincer, G. Akpolat, and E. Zeydan, "Security issues of big data applications served by mobile operators," in *Proc. 25th Signal Process. Commun. Appl. Conf.*, 2017, pp. 1–4.

[129] S. Yakoubov, V. Gadepally, N. Schear, E. Shen, and A. Yerukhimovich, "A survey of cryptographic approaches to securing big-data analytics in the cloud," in *Proc. IEEE High Perform. Extreme Comput. Conf.*, 2014, pp. 1–6.

[130] L. Dupré and Y. Demchenko, "Impact of information security measures on the velocity of big data infrastructures," in *Proc. Int. Conf. High Perform. Comput. Simul.*, 2016, pp. 492–500.

[131] N. Chaudhari and S. Srivastava, "Big data security issues and challenges," in *Proc. Int. Conf. Comput. Commun. Autom.*, 2016, pp. 60–64.

[132] M. Paryasto, A. Alamsyah, B. Rahardjo, et al., "Big-data security management issues," in *Proc. 2nd Int. Conf. Inf. Commun. Technol.*, 2014, pp. 59–63.

[133] R. Clarke, "Quality assurance for security applications of big data," in *Proc. Eur. Intell. Secur. Informat. Conf.*, 2016, pp. 1–8.

[134] I. Škrjanc, A. S. de Miguel, J. A. Iglesias, A. Ledezma, and D. Dovžan, "Evolving cauchy possibilistic clustering based on cosine similarity for monitoring cyber systems," in *Proc. Evolving Adaptive Intell. Syst.*, 2017, pp. 1–5.

[135] S. Alouneh, I. Hababeh, F. Al-Hawari, and T. Alajrami, "Innovative methodology for elevating big data analysis and security," in *Proc. 2nd Int. Conf. Open Source Softw. Comput.*, 2016, pp. 1–5.

[136] K. Dhanasekaran and B. Surendiran, "Nature-inspired classification for mining social space information: National security intelligence and big data perspective," in *Proc. Online Int. Conf. Green Eng. Technol.*, 2016, pp. 1–6.

[137] K. D. Strang and Z. Sun, "Meta-analysis of big data security and privacy: Scholarly literature gaps," in *Proc. IEEE Int. Conf. Big Data.*, 2016, pp. 4035–4037.

[138] L. Yuqing, "Research on personal information security on social network in big data era," in *Proc. Int. Conf. Smart Grid Electr. Autom.*, 2017, pp. 676–678.

[139] J. Hariharakrishnan, S. Mohanavalli, K. S. Kumar, et al., "Survey of pre-processing techniques for mining big data," in *Proc. Int. Conf. Comput. Commun. Signal Process.*, 2017, pp. 1–5.

[140] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Commun. Mag.*, vol. 53, no. 10, pp. 190–199, Oct. 2015.

[141] T. Lu, X. Guo, B. Xu, L. Zhao, Y. Peng, and H. Yang, "Next big thing in big data: the security of the ict supply chain," in *Proc. Int. Conf. Social Comput.*, 2013, pp. 1066–1073.

[142] E. Damiani, C. Ardagna, F. Zavatarelli, E. Rekleitis, and L. Marinos, "Big Data Threat Landscape," *Eur. Union Agency Netw. Inf. Secur.*, Jan. 2017. [Online]. Available: https://www.enisa.europa.eu/publications/bigdata-threat-landscape

[143] R. Geambasu, T. Kohno, A. A. Levy, and H. M. Levy, "Vanish: Increasing data privacy with self-destructing data," in *Proc. USENIX Secur. Symp.*, 2009, vol. 316.