Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2020, 7(9):85-93

Research Article ISSN: 2394 - 658X

An In-Depth Exploration of Techniques for Utilizing Big Data in Network Security Analytics

Kartheek Pamarthi

Kartheek.pamarthi@gmail.com

ABSTRACT

The ability to process huge amounts of complicated data enables businesses to discover previously unseen patterns, get insights, and even exchange data over the network. A platform that is both efficient and safe is required since the data that is transferred across the network of an organisation is frequently sensitive. Because of this, network security has been elevated to a position of prominence in the era of big data. Within this framework, network security platforms are required to manage huge amounts of complicated information in order to anticipate and thwart potential threats in real time. The fact that these platforms are frequently based on conventional methods, on the other hand, renders them unreliable for the protection of large amounts of data. Throughout the course of this article, we will primarily concentrate on the protection solutions for large data and network security. The first thing that we will do is discuss the elements that influence network security platforms in the era of big data. Then, while conducting a study of recent research, we go over a variety of big data solutions that enable the protection of distributed networks.

Keywords: Network Security, Big data, Fault detection.

INTRODUCTION

According to [1], "measures taken to protect a computer or computer system (as on the Internet) against unauthorised access or attack" is the definition of computer security. "Strategy, policy, and standards regarding the security of and operations in cyberspace, and encompass[ing] the full range of threat reduction, vulnerability reduction, deterrence, international engagement, incident response, resilience, and recovery policies and activities, including computer network operations, information assurance, law enforcement, diplomacy, military, and intelligence missions as they relate to the security and stability of the global information and communications infrastructure" is another long definition of cyber-security offered by the US national initiative for cyber-security careers and studies (NICCS). More recently, in 2015, 1,966,324 alerts were recorded regarding malware infections that attempted to infiltrate online bank accounts in order to steal money, as seen in the Kaspersky final statistic report [2].

The enormous threat that malware infections pose to the global economy each year is highlighted by this staggering number of infections in just one economic sector. Given the exponential growth of cyberattacks, it is evident that current measures to protect IT infrastructure, company networks, and online applications may be insufficient. Therefore, one might wonder: what can be done to better safeguard against and identify the exponential rise of these cyberattacks? A 2015 survey by International Data Corporation (IDC) polled security professionals, executives, and experts from the US government, energy sector, and financial services sector. To have a better grasp of how cyber risks have changed over time, we conducted this interview. Respondents to the interviews agreed that cyber-security risks are on the rise and that businesses should stop responding to security incidents after the fact and start proactively assessing potential dangers [3].

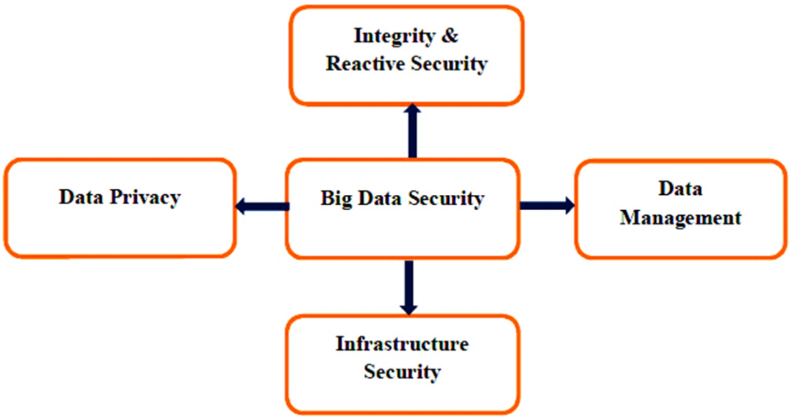

Figure 1. Security of Big Data

Security measures are those put in place to prevent unauthorised individuals from gaining access to, altering, disrupting, inspecting, disclosing, recording, or destroying data or information assets. Data and information assets are safeguarded in these drills by means of training, procedures, and technology. Data protection from hostile assaults and the improper exploitation of stolen data for financial benefit is a primary focus of the security approach. The challenges of securing large data can be classified into four main areas, as illustrated in Figure 1: data management, infrastructure security, data privacy measures, and integrity and reactive security. Nonrelational data stores and secure distributed programming should make up the infrastructure's security measures.

Subsequently, "data privacy" is defined as the safeguarding of individual records by means of granular access control and encrypted data centres.

Depending on their size, huge businesses are responsible for producing up to one hundred billion events every single day, as stated in [4]. A number of factors, including the expansion of data sources, the number of personnel, the introduction of new equipment, and the execution of additional software programmes, can all have an impact on the number of events that occur. Existing cybersecurity analytics techniques, such as intrusion detection systems and log event analysis, are inadequate, scale poorly, and frequently produce a high number of false positives. Organisational adoption of cloud architectures and data accumulation exacerbate the problem [4]. Additionally, cyber threat intelligence is essential for ensuring continuous real-time data gathering and monitoring, which in turn allows for the right risk mitigation process to be begun before attacks can cause significant damage.

SIEM, which stands for security information and event management, is a platform that is used for the purpose of gathering and correlating data on network flow, logs, and security events for the purpose of conducting security operations and conduct security analysis [5]. The term "cyber threat intelligence" refers to a set of capabilities included in SIEM systems, which help with things like log management, security event correlation, and monitoring network activities [6]. The majority of companies are now using security information and event management (SIEM) systems to monitor possible threats rather than traditional security analysis and investigations [5]. However, there are a lot of issues with the Security Information and Event Management (SIEM) system that make it unable to handle the massive change in modern threats.

For example, according to [5], the following are examples of these deficiencies: New attacks that use multidimensional tactics and differ from system to system are difficult to react to because event correlation in SIEM relies on data that has been normalised in reference to established schemas. Second, because platforms that implement SIEM are dependent on fixed storage (Schema), it becomes challenging to deal with the ever-increasing number of events that are generated over the course of time for these platforms; Third: Security information and event management systems are based on a preset context, which means that they are context-specific and cannot be generalised to various scenarios unless they are specified again, which is a process that also takes a lot of time; Four, security information and event management systems are so rigid that the addition of new regulations necessitated the complete rethinking of the strategy. According to the findings presented in [6], the authors came to the conclusion that businesses require a new strategy for cyber-security that is capable of overcoming the deficiencies outlined earlier. According to reports, the new method may be understood as an end-to-end link between security analytics tools that are based on Big Data and the expertise of the information security team.

LITERATURE REVIEW

“Big Data” is a relatively new phrase used to describe datasets that have beyond the capabilities of traditional database management systems. Massively large data sets that aren’t amenable to typical data storage and management tools or methods [7].

These technologies and systems aren’t up to the task of dealing with their magnitude. Sizes of big data are always increasing; currently, a single data set can contain information ranging from several petabytes (PB) to a few dozen terabytes (TB). The difficulty in capturing, storing, searching, sharing, analysing, and visualising huge data is one of the problems that stem from this. Current commercial practices involve sifting through extremely detailed data sets in an effort to find insights that were previously hidden [8].

Characteristics Of Big Data

Obtaining insights that show new avenues to enterprise value from data that is vast, scattered, varied, and/or time-sensitive requires new technical architectures, analytics, and tools. Big data is defined by three characteristics: velocity, diversity, and volume. The data's breadth and magnitude determine its volume. To put it simply, "velocity" is the rate of data production or progression. Lastly, diversity encompasses not only the various data types and formats, but also their applications and analyses [9]. The sheer volume of data is the defining characteristic of big data. One way to quantify large data is by looking at the amount of records, transactions, tables, or files. Another option is to consider its size in terabytes or petabytes. Another important consideration is the proliferation of sources for big data, which now includes social media, clickstreams, and logs among others. Incorporating varied sources into analytics brings together a variety of data types, including text, human language, and semi- structured data like eXtensible Markup Language or RSS feeds, as well as traditional structured data. The fact that data comes from many places further complicates its categorization. This class includes media players (both audio and video). Additional historical context can be added to big data by retrieving multi-dimensional data from a data warehouse. The crux of big data is variety, not just quantity.

The speed of big data is another useful metric. Here we get the fundamental data transfer rate. As far as large data types go, streaming data—information retrieved in real-time from websites—is considered state-of-the-art [10]. The possibility of a fourth letter for "veracity" has been considered by a number of forums and specialists. Paying close attention to the data quality is the key to accuracy. This ranks the quality of big data as great, terrible, or undefinable because of issues such data inconsistency, incompleteness, ambiguity, latency, dishonesty, and approximations [11].

Big Data Analytics Tools and Methods

There is a growing need for more efficient and quicker methods of data analysis due to the exponential growth of both technology and the quantity of data flowing into and out of businesses on a daily basis. You can't just have a plethora of facts and expect to make fast, effective decisions anymore. Traditional methods of data management and processing are inadequate when faced with such massive data collections. This highlights the importance of developing new infrastructures for storing and handling massive amounts of data, in addition to new tools and approaches for big data analytics. All steps of the data lifecycle, from gathering raw data to cleaning it up and analysing it, are impacted by the advent of big data.

This led to the proposal of the B-DAD paradigm by [12], which integrates decision-making with big data analytics frameworks and techniques. Each step of the decision-making process is accompanied by related activities, such as tools for analytics, visualisation, and evaluation; tools for large data storage, administration, and processing; and so on. Data and analytics processing, big data storage and architecture, and big data analytics that can lead to better decision-making and the discovery of new information are the three primary sectors that have been greatly affected by big data analytics. In what follows, we'll delve deeper into each of these topics.

The goal of this section is to provide a high-level overview, not a comprehensive rundown of all potential opportunities and technology.

Due to the ever-changing nature of big data as a study subject and the ongoing development of new results and tools, it is not exhaustive.

Big Data Storage and Management

Deciding where and how to store the data after collection is one of the first tasks for firms dealing with big data. Data warehouses, data marts, and relational databases are some of the most traditional ways to store and retrieve structured data. Using tools that modify data from external sources to meet operational needs, operational data stores can be imported into databases and data warehouses. Extract, Modify, Load, or ETL, is an acronym for these. Data cleansing, transformation, and categorization are further steps in getting the data ready for data mining and other forms of internet analytics (12).

Big data environments differ from traditional Enterprise Data Warehouses (EDWs) in that they call for the utilisation of MADD (Magnetic, Agile, and Deep) analysis techniques. Before adding new data sources to the system, typical EDW approaches recommend cleaning and adding them. Big data environments need to be magnetic if they are to accept any kind of data source [14]. This is due to the fact that data is ubiquitous in today's world. Because data sources are multiplying and data studies are becoming more complicated, it is essential that analysts be able to rapidly create and change data through big data storage.

In order to keep up with the ever-changing data landscape, a database with adaptable logical and physical contents is essential [15]. In addition to storing massive amounts of data, a big data repository should act as an advanced algorithmic runtime engine [16]. This is because modern data studies employ intricate statistical approaches, and analysts must possess the ability to drill down into massive datasets in order to draw meaningful conclusions. This led to the implementation of numerous solutions for big data. These solutions can take many forms, including distributed systems, databases constructed with Massive Parallel Processing (MPP) for both high query performance and platform scalability, non-relational databases, and databases in memory. Databases that are not relational, such as Not Only SQL (NoSQL), were created to facilitate the storage and management of unstructured data, which is not classified as relational. The purpose of NoSQL databases is to facilitate application development and deployment with less effort, provide data model design freedom, and expand to large scales. Unlike relational databases, which merge data storage and management, NoSQL databases keep these two functions distinct. These databases enable data administration activities to be done at the application layer instead of in database-specific languages, and their focus is on scalable and high-performance data storage [17]. On the other hand, in-memory databases eliminate disc I/O and provide immediate results by storing all data directly in the server's RAM.

Using silicon-based primary memory instead of mechanical disc drives to hold the major database is viable. There are several benefits to this. Because of this, not only is it feasible to create entirely new apps, but performance is also enhanced by a factor of ten higher than before. Plus, in-memory databases are already in use for sophisticated analytics on massive datasets, mainly to speed up the scoring and access to analytical models. This provides speed for discovery analytics and scalability for massive data sets [18].

Big data analytics, on the other hand, are made possible by Hadoop, a framework. It provides dependability, scalability, and management by offering an implementation for the MapReduce paradigm, which will be covered more below. Hadoop also handles the integration of analytics and storage. Hadoop is composed of two primary components: MapReduce, which performs analysis on the same enormous data sets, and HDFS, which stores the data sets [19].

A dependable and redundant distributed file system is provided by the HDFS storage capability for large files. One way that HDFS's storage function helps with large files is by dividing them into blocks and then distributing those blocks among cluster nodes. Data security among nodes is further enhanced by utilising a replication method. This approach guarantees that the data will remain accessible and trustworthy even if a node fails. Data nodes and name nodes are the two main categories of HDFS nodes. In order to facilitate communication between the client and the Data Node, the Name Node mediates between the two parties [20]. Distributed across a large number of Data Nodes are duplicated file blocks that hold the data.

THE ELEVEN VS (11VS) OF BIG DATA

The three hallmarks of large data stores and databases in the past were diversity, velocity, and volume. Subsequently, the phrase “big data” was used to describe the enormous volumes of data available in the contemporary digital age. On the other hand, new technological developments have given big data additional characteristics that are increasingly relevant to its semantics. Credibility, or the quality of facts, is one of the most important extra factors that have been acknowledged. Consequently, the three most important aspects of large data are validity, volatility, and value. Additionally, there have been developments in some aspects of large data, such as its visualisation and its variability, and an increasing number of technical difficulties are being recognised within big data. Modern big data presents unique privacy and security concerns, which have been linked to the Valence and Vulnerability dimensions. This section aims to prepare us for the security concerns that will accompany future big data endeavours by describing the qualities and characteristics of all eleven dimensions of big data, which are expanding beyond the traditional 3Vs. A typical big data system identifies the 11Vs as dimensions (see Figure 2). Detailed descriptions of these 11Vs follow.

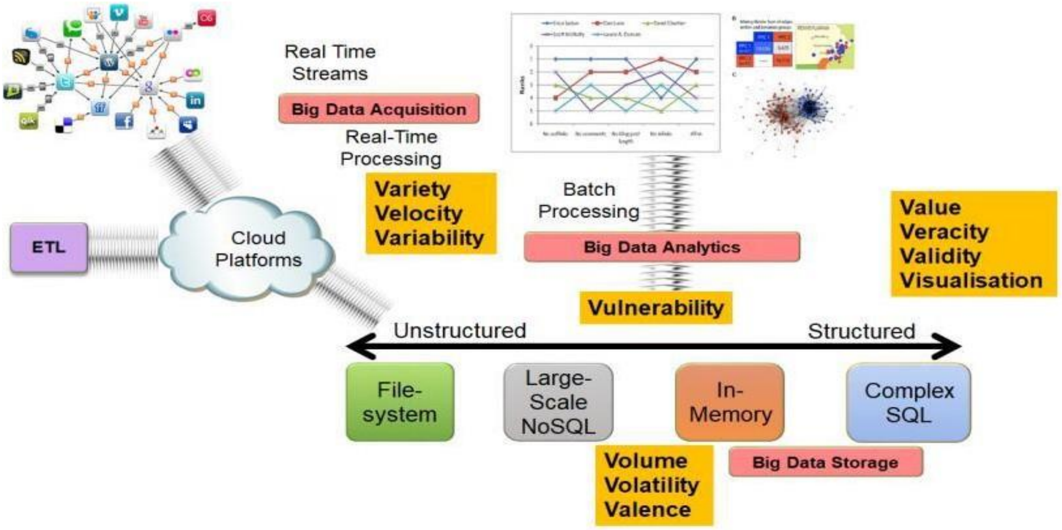

Figure 2. The typical dimensions of a big data system.

Dimension 1: Volume

The first known attribute of big data is volume, which is the overall quantity of data acquired. Modern data has been largely collected in the past few years, thanks to the convergence of commercial websites, social engineering, and mobile applications (Apps).

Data generation and storage are increasing at an exponential rate every day. The number of large multimedia files uploaded to social media sites like YouTube and others every minute is at least 300 hours of video. Businesses are storing several billions of records due to the combination of traditional transactional data with data shared on social media platforms. The capacity of a single data source could be rising at a rate ranging from petabytes to zettabytes. Data on a worldwide scale will reach 175 zettabytes by 2025, up 61% from now, says a new report from International Data Corporation (IDC).

Big data's sheer quantity affects privacy and security in two main ways, which are detailed below:

i. The inability of conventional database management systems and software tools to continually oversee and implement uniform security standards is caused by the dispersed nature of data storage across different nodes, servers, and clusters.

ii. Data transactions and performance under tolerance time limits are susceptible to security vulnerabilities and the possible effects of cluster or node failures in this configuration.

Dimension 2: Velocity

Velocity, the second dimension, is concerned with the velocity of data generation and inflow into organisations, as well as the ever-increasing demand for real-time processing of this data. This has an effect on big data analytics, which rely on computers' ability to handle large amounts of data quickly enough in real time. Although there is room for improvement in data storage capacity, the rate of data generation should be prioritised. Business prospects may still go unrealized if data cannot be analysed in real-time, regardless of how readily available it is. For example, if the processing speed of weather predictions is too sluggish to keep up with the speed of data received, it will impact the ability to make timely judgements accurately. The security and privacy of huge data sets are impacted by the velocity of data because real-time transactions require faster cryptographic algorithms. Privacy policies must also be up-to-date to accommodate the rapid data buildup, which means that security audits must be conducted to record and monitor past data.

Dimension 3: Variety

The variety dimension offers a potential framework for describing the diversity of organised, semi- structured, and unstructured large data. Files containing audio, video, and sensor signals make up the bulk of the unstructured data, along with logs from various equipment, social media, and satellites. Different data representations are just one aspect of the Variety dimension; it also includes the ways and modalities of conveying the same information. Along with the structural diversity shown by the most frequent variety, there is also media variety and semantic variety. Media variety relates to the numerous ways in which the same data is presented, while semantic variety shows that different data settings generate different interpretations. Unstructured data lacks latent meaning, however structured data can have its semantic meaning communicated using a typical query in structured query language (SQL). Many forms of semi- structured data have emerged as a result of the widespread use of markup languages like XML and email in recent years. With so many different kinds, formats, and sources of big data, it's important to properly categorise and restrict access to data in order to protect user privacy and security.

Dimension 4: Veracity

Another important aspect of big data that has recently come to light is veracity, which pertains to the authenticity or quality of the data. This is in response to the increasing concerns around data availability and streaming. To make big data work with useful analysis, you need the appropriate data, enough of it, and processed when it's needed. Information that is duplicated, missing, or inaccurate will not yield useful results when applied to data analysis.

There is less trust and confidence in the data as the veracity decreases as the first three Vs rise with big data. The decision-making risks faced by businesses may be mitigated if big data were more reliable. When it comes to implementing data ownership and periodic access review processes to ensure high- quality data, this affects privacy and security regulations.

Dimension 5: Validity

The fifth component, Validity, is also associated with Veracity; it describes how well data fits a certain context or is used for its intended purpose. Therefore, in order to profit from big data analytics in context, validity ensures that the data is correct for a certain application or perspective of the data. Maintaining the veracity of the data cleansing process—which many businesses invest significant time in before doing data analysis—requires robust data governance procedures. The entire data supply chain must be protected, which necessitates the correct management of third-party vendors and partners.

Dimension 6: Volatility

Big data quality assurance is defined by validity and veracity, but there is also a temporal dimension to data called volatility that dictates how long data is legitimate to be stored. In order to do real-time analysis, this metric ensures that all datasets are current. The costs and storage constraints of cloud computing make it imperative to have stringent backup and archiving strategies in place before deciding how long data should be retained. Regular archiving of obsolete and unimportant data is necessary to enhance the efficiency of big data analytics. Big data's unpredictability affects privacy and security regulations, as well as processes for data deletion, periodic re-evaluation of security solutions, and data retention.

Dimension 7: Value

The seventh dimension of big data has been established, which helps to understand the advantages of big data as they relate to various stakeholders inside an organisation who add value to their business. Several elements that address concerns like: when is the best time to make judgements, which business decisions could benefit from big data insights, and who would directly reap the benefits are all part of this value dimension. Value, in a word, is the measurement of big data's utility in decision-making for the purpose of enhancing company performance. In order to solve their complicated business problems, businesses need to get data insights, and this is where big data analytics come in. The right big data strategy can then be launched with its assistance. Since analytics serve as a springboard for action in businesses, it is essential to have appropriate authorization and access controls over analytical assessments. Big data also has its own significance in this regard. The establishment of suitable security gates throughout the creation of such data insights is equally crucial.

Dimension 8: Variability

The eighth dimension of big data, variability, could impact all seven of the previously mentioned Vs. Variations in data loading speeds, formats, or types may arise from inconsistencies in the manner data is loaded into storage from various sources, which is described by this dimension. Another benefit to businesses is the ability to detect anomalies and outliers. Recording and linking such information about the data's variability with the data stored in the data warehouse is important to generate relevant insights from big data. The IT security operations should modify their various audit log gathering and monitoring strategies to account for the substantial data fluctuation.

Dimension 9: Visualisation

The visualisation of big data has recently grown in importance as a means to better understand the data. Data visualisation methods like k-means clustering, heat maps, cone trees, and dashboards are all part of this category. Data visualisation simplifies its display, and conventional methods of data communication relied on basic charts and graphs. More relevant graphical interpretations of big data are being made possible through the integration of numerous advanced visualisation tools with data analytics models. This is all in an effort to aid in successful decision making. As a result, we think of visualisation as a common need that makes up the ninth dimension of big data. Not only should access controls and privileges be assigned according to user roles and responsibilities, but privacy and protection policies should also be set up for the outputs of different visualisation tools.

Dimension 10: Valence

Massive amounts of data won't help with big data if we can't find a way to link the many pieces of information together. Then we'll just have islands of unrelated data, with no idea how they all fit together. Data captured while streaming could lead to the establishment of any direct connections. Indirect relationships between data items are more valuable to the company, but they are also more difficult to detect. This web of relationships gives rise to the tenth dimension, or valence, of big data, much like the bonds between atoms in a molecule. Counting the number of real connections relative to the total number of potential connections in the dataset is one approach to measure data density. This ratio is called Valence. To keep the right valence that supports the big data system's service level agreement (SLA), scalability of hardware, networks, and systems is necessary, especially given the high number of heterogeneous data access points. Big data systems should be able to continue performing at their current level as they expand in the future, thanks to security and privacy management methods.

Dimension 11: Vulnerability

Big data's Vulnerability property is the last and most crucial feature; it concerns the technological, privacy, and security dangers associated with the many forms of rich personal data acquired through products and services through online apps, social media, and Internet of Things (IoT) devices. The management, procedures, and technology surrounding big data are susceptible to security and privacy breaches caused by a lack of standards. As a result of allegations of security and privacy breaches, the Vulnerability dimension of big data has lately garnered a lot of attention. Rules and procedures for big data incident management for security and privacy are based on the aforementioned areas that require ongoing vigilance. Regular vulnerability evaluations and penetration testing should be developed to accommodate the unique features of big data. Finding potential entry points that could compromise the privacy, security, and accessibility of big data systems and data is crucial for preventing the leakage of sensitive information.

SECURITY AND PRIVACY CHALLENGES OF BIG DATA

The massive increase in data in the modern day has led many to turn to cloud storage services. Big data mining makes more data available and helps organisations learn a lot, which in turn helps them make better business decisions. Big data has the potential to improve people's and companies' day-to-day lives in numerous ways, but its usage in intelligent services poses risks to people's privacy and security. Cybercriminals consider innovative and user-friendly goods and services like the Internet, the Internet of Things (IoT), smart devices, and social networks as an opportunity to profit from their bad activities.

Such sensitive and important data could be released, capturing rich information on a person's interests, preferences, movements, and behaviour patterns. That is why, all through big data's lifespan, proper privacy and security measures must be implemented. At each step of the big data lifecycle—data collection, storage, and analysis—we detail the security and privacy concerns. Below is Figure 3, which summarises the three stages and their effects on big data security and privacy.

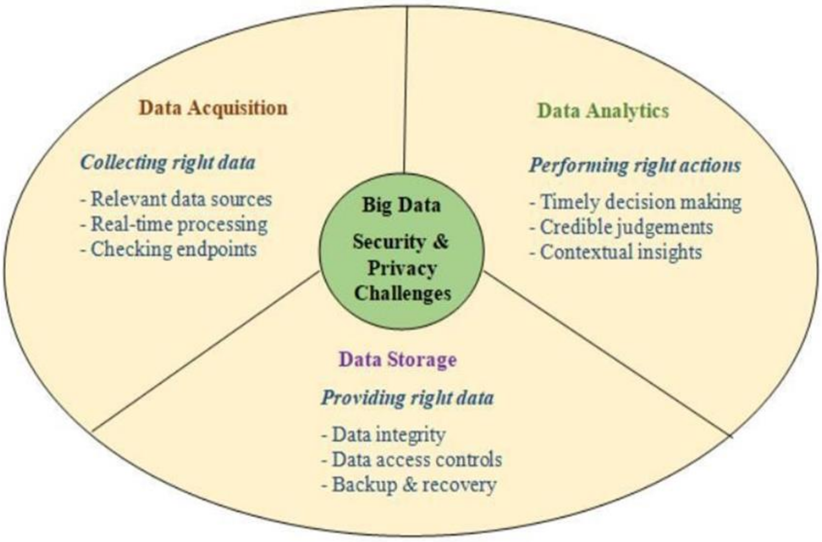

Figure 3. Concerns about data privacy and security in the age of big data.

Big Data Acquisition

Big data gathers a wide variety of data from different online services and commercial transactions, including structured, unstructured, semi-structured, and real-time processed data. Here are some of the most pressing issues we've seen so far, and how they relate to the four main pillars of big data:

- The diversity of big data means that traditional encryption-based security solutions aren't going to cut it for every source.
- Discrepancies in data formats, speeds, and types, such as web clickstream data, might cause security risks since big data is unpredictable.
- Real-time traffic monitoring is difficult due to the speed of data acquisition.

Taken together, these concerns mean that acquired big data has the potential to harbour advanced persistent threats (APTs). APT is most effective in social networks where data streams originate from a wide variety of sources and use non-standard formats. It becomes more challenging to detect APT code in real-time when it is concealed in huge data. Attackers may build a botnet that targets the data source, destination, and all connectivity by taking advantage of their vulnerabilities. The integrity of big data in a real-time processing environment depends on data security and privacy standards that are implemented from the very beginning of the data collecting process. Connecting the right network endpoints for data flow is essential for big data, but so are sophisticated authentication and privacy requirements.

Big Data Storage

At this stage of the big data life cycle, the data that has been gathered from various sources is stored and organised. A big data storage system's principal objective is to make it easier for the company's internal and external stakeholders to retrieve any relevant information whenever they need it in the future.

The following are the most important difficulties of this stage:

- With the ever-increasing size of big data, the volume dimension affects the server architecture of a company. Traditional data warehouses might not be able to cope with the ever-increasing pace and volume of big data, hence it might be essential to use distributed, cloud, or other outsourced big data servers.
- A large amount of data, including structured and unstructured information, as well as semi-structured data, is being stored in big data systems. Online purchases, consumer reviews, social media posts, emails, marketing records, and other logs pertaining to a business's operations are among the many sources of this data.
- For the purpose of their functional operations and day-to-day transactions, other connected departments also share big data. Thus, the density of big data, which is a result of the connectivity of multiple in-house, cloud-based, or outsourced data centres, might impact the value of big data.

For the reasons stated above, increasing the real-time speed of multi-party activities on the same data storage could compromise the data's integrity. Due to the possibility of various processes, traditional security solutions such as encryption may not be effective in maintaining data integrity. Sniffers could be more likely to access the servers in such a disjointed setting by taking advantage of security policy gaps and inconsistencies. A breach of privacy could occur as a result of data misuse. Theft of sensitive user data and invasion of privacy are made more likely by these reasons. No amount of standard role-based, mandatory, or discretionary access control has been found to be effective in a big data storage environment due to the large diversity of users and their varying degrees of permission. This emphasises the significance of having up-to-date, trustworthy data access rules that adhere to all applicable security and privacy laws. Good backup and recovery procedures must be followed at all stages of the big data life cycle when dealing with data that is either outdated and needs to be removed or archived.

Big Data Analytics

Collecting and storing big data using quality data storage systems allows us to evaluate, analyse, and draw insights from the data, which in turn allows us to make better, more timely decisions. Accordingly, big data analytics—a crucial step in the big data life cycle—is the principal objective. For the purpose of developing a more effective marketing plan to offer customers individualised products and better services, businesses gather a great deal of sensitive and contextual data about their customers through analysis of their interactions. There may be unanticipated privacy breaches as a result of cloud-based analysis of such data from several sources. Data analytics usually consists of three primary steps: (i) preparing the data by identifying, cleaning, and formatting it to meet the analytics model's needs; (ii) adopting the model; and (iii) communicating the output to give insights derived from the data. Due to their inherent vulnerabilities, each of these procedures encounters several security and privacy concerns. Here are four major obstacles we found throughout the big data analytics stage:

I. Big data analytics yield value depending on the reliability of the data. Analytics based on data could be misleading if it originates from a shady source or has been altered in transit.

II. Users' privacy can still be compromised by big data analytics, even after anonymizing the data. The data analytics' trustworthiness is called into doubt when the Veracity component of big data is compromised.

III. When applied to a particular context, big data analytics use cognitive algorithms and machine learning to sift through massive amounts of data. Hacked data might compromise this validity aspect of big data. One example is the development of new apps that can filter emails based on subject or if they are spam. The end result of this procedure is a dictionary of officially recognised words. In order to steal sensitive information and contextual data from big data, hackers often use these apps as a cover to hide malicious code.

CONCLUDING REMARKS

The dynamic integration of several technologies through Intranets, cloud infrastructures, the Internet, social media, and the Internet of Things (IoT) networks has led to the tremendous complexity and diversity of big data systems. This essay made an attempt to comprehend the big data system holistically by using first principles thinking, with the goal of addressing the increasing amount of privacy and security problems. Before diving into the intricacies of the big data environment, we set out to grasp its foundation in order to tackle these difficulties. The eleven crucial features of big data—volume, velocity, variety, veracity, validity, volatility, value, variability, visualisation, valence, and vulnerability—are directly or indirectly affecting the increasing privacy and security concerns. These dimensions are a product of how various parts of big data have developed over time. Secondly, we considered the three main phases of a big data system's life cycle: data collection, storage, and analysis. The 11Vs of big data were also assigned to each step, taking security and privacy into consideration. Finally, we covered all the bases in our practical four-solution approach to the privacy and security concerns around big data. These tactics make use of particular pieces of contemporary technology. In order to tackle the security issues that arise during the big data life cycle, four current technologies were reviewed, each with their own unique adaptation: blockchain, data mining, data provenance, and data encryption and access control. In conclusion, this article uncovered some obstacles, unanswered questions, and potential approaches to implementing technology in the field of big data security. Future research on this crucial topic may be prompted by the underlying implications.

REFERENCES

[1]. 2019 cyber security statistics trends & data: The ultimate list of cyber security stats — purplesec. https://purplesec.us/resources/cyber-security-statistics/. (Accessed on 07/30/2020)

[2]. 2020 trustwave global security report — trustwave. https://www.trustwave.com/enus/resources/library/documents/2020-trustwave-global-security-report/. (Accessed on 08/01/2020)

[3]. 5 cybersecurity threats to be aware of in 2020 — ieee computer society. https://www.computer.org/publications/tech-news/trends/5-cybersecuritythreats-to-be-aware-of-in- 2020/. (Accessed on 07/30/2020)

[4]. Apple reveals windows 10 is four times more popular than the mac. https://www.theverge.com/2017/4/4/15176766/apple-microsoft-windows-10- vs-mac-users-figures- stats. Accessed: 2018-12-03

[5]. Computer science. https://arxiv.org/archive/cs. (Accessed on 07/30/2020)

[6]. Cyberthreat trends: 15 cybersecurity threats for 2020 — nortonlifelock. https://us.norton.com/internetsecurity-emerging-threats-cyberthreattrends-cybersecurity-threat- review.html. (Accessed on 07/30/2020)

[7]. Github - mozilla/openwpm: A web privacy measurement framework. https://github. com/mozilla/OpenWPM. (Accessed on 03/23/2019)

[8]. Global 2020 forecast highlights. https://www.cisco.com/c/dam/m/en_us/solutions/ service- provider/vni-forecast-highlights/pdf/Global_2020_Forecast_ Highlights.pdf. (Accessed on 07/30/2020)

[9]. Half of the malware detected in 2019 was classified as zero-day threats, making it the most common malware to date - cynet. https://www.cynet.com/blog/half-ofthe-malware-detected-in-2019-was- classified-as-zero-day-threats-makingit-the-most-common-malware-to-date/. (Accessed on 07/30/2020)

[10]. Microsoft vulnerabilities more than doubled in 2017. https://www.securitynow.com/ author.asp?section_id=649&doc_id=740671. Accessed: 2018-12-03

[11]. Top cybersecurity threats in 2020. https://onlinedegrees.sandiego.edu/topcyber-security-threats/. (Accessed on 07/30/2020)

[12]. Ransomware cyber attacks: Which industries are being hit the hardest? https://www.bitsighttech.com/blog/ransomware-cyber-attacks (2017). (Accessed on 12/08/2018)

[13]. Us hospital pays $55,000 to hackers after ransomware attack — zdnet. https://www.zdnet.com/article/us-hospital-pays-55000-to-ransomware-operators/ (2018). (Accessed on 12/08/2018)

[14]. Abdlhamed, M., Kifayat, K., Shi, Q., Hurst, W.: Intrusion prediction systems. In: Information Fusion for Cyber-Security Analytics, pp. 155–174. Springer (2017)

[15]. Abraham, S., Nair, S.: Predictive cyber-security analytics framework: a nonhomogenous markov model for security quantification. arXiv preprint arXiv:1501.01901 (2015)

[16]. Aditham, S., Ranganathan, N.: A system architecture for the detection of insider attacks in big data systems. IEEE Transactions on Dependable and Secure Computing 15(6), 974–987 (2018)

[17]. Alani, M.M.: What is the cloud? In: Elements of Cloud Computing Security, pp. 1–14. Springer (2016)

[18]. AlEroud, A., Karabatis, G.: Using contextual information to identify cyber-attacks. In: Information Fusion for Cyber-Security Analytics, pp. 1–16. Springer (2017)

[19]. Aleroud, A., Zhou, L.: Phishing environments, techniques, and countermeasures: A survey. Computers & Security 68, 160–196 (2017)

[20]. Alguliyev, R., Imamverdiyev, Y.: Big data: big promises for information security. In: Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on, pp. 1–4. IEEE (2014)