AIMS Mathematics, 4(3): 860–879.
DOI: 10.3934/math.2019.3.860
Received: 25 April 2019
Accepted: 01 July 2019
Published: 19 July 2019

http://www.aimspress.com/journal/Math

Review

Big data security challenges and strategies

Sitalakshmi Venkatraman[1,*] and Ramanathan Venkatraman[2]

[1] Department of Information Technology, Melbourne Polytechnic, VIC, Australia
[2] Institute of Systems Science, National University of Singapore, Singapore

*** Correspondence:** Email: sitavenkat@melbournepolytechnic.edu.au; Tel: +61892663143.

Abstract: Big data, a recently popular term that refers to a massive collection of very large and complex data sets, is facing serious security and privacy challenges. Due to the typical characteristics of big data, namely velocity, volume and variety associated with large-scale cloud infrastructures and the Internet of Things (IoT), traditional security and privacy mechanisms are inadequate and unable to cope with the rapid data explosion in such a complex distributed computing environment. With big data analytics being widely used by businesses and government for decision making, security risk mitigation plays an important role in big data infrastructures worldwide. Traditional security mechanisms have failed to cope with the scalability, interoperability and adaptability of contemporary technologies that are required for big data. This paper takes an exploratory initial step using first principles to address this gap in literature. Firstly, we establish the current trends in big data comprehensively by identifying eleven Vs as important dimensions of big data, which form the contributing factors having an impact on the impending security problem. Next, we map the eleven Vs to the three phases of big data life cycle in order to unearth the major security and privacy challenges of big data. Finally, the paper provides four practical strategies adapted from contemporary technologies such as data provenance, encryption and access control, data mining and blockchain, identifying their associated real implementation examples. This work would pave way for future research investigations in this important big data security arena.

Keywords: big data; security risks; privacy issues; cloud computing; data analytics; blockchain technology

1. Introduction

With the application of Internet and mobile technologies prevalent in everyday life, including

social networks, Internet of Things (IoT) and personalised services, huge amounts of data are continuously being collected, stored, analysed and utilised in various platforms including the cloud by individuals and organisations [1,2]. Big data, a term used for such a collection of massive datasets, possesses typical characteristics such as, fast-moving, multi-source origin, tremendously large and unstructured [3,4]. These features define the three well-known dimensions of big data, namely velocity, variety and volume, which are referred to as 3Vs [5]. Big data is produced from different websites, multimedia archives, social networks and IoT networks that connect a variety of devices and sensors. Recently, big data has become a hot topic with significant impact, transforming industries worldwide. Businesses and government organisations consider big data analytics as a contemporary and valuable technique to analyse complex and historical data to discover patterns that could support in their effective decision-making. Big data plays an important role in future data management and operations in various industry sectors such as healthcare, manufacturing, retail, traffic management, banking, weather bureau, education and transportation [6,7]. Many research studies have unearthed several benefits of big data applications. However, recent literature surveys conducted in the topic of big data security indicate that malicious attackers targeting big data have been on the rise [8]. The prime issues and solutions around the security risks and privacy protection are still not explored comprehensively in the big data domain [9,10]. These challenges motivate new innovations and research activities to discover open issues that pave the way for future research and practice [11,12]. In line with this objective, this paper first examines the various characteristics of big data that are contemporary to the current trends, with the aim to provide a comprehensive overview of its fundamental concepts. This has resulted in identifying additional dimensions of big data resulting in eleven Vs (11Vs). Another unique contribution of the paper is that the 11Vs are viewed in relation to the security and privacy issues and threats that have been identified during every phase of the big data life cycle, right from data acquisition phase to data storage phase and finally into the data analytics phase. The third distinct contribution is the proposal of four practical strategies as counter measures to effectively address the threats, risks and vulnerabilities of big data along with real examples cited from literature.

The rest of the paper is organised as follows. Section 2 provides a literature review of the related work and identifies the need for the study. Section 3 describes the evolution of various dimensions of big data that lead towards identifying eleven dimensions of today's big data environment. In Section 4, we comprehensively describe the security and privacy issues of big data during the three phases of its life cycle, mapping them to the 11Vs. In Section 5, we propose four strategies to overcome big data security challenges based on contemporary and futuristic technologies. Finally, concluding remarks are provided in Section 6.

2. Literature review

During the last decade, big data has been increasingly adopted in almost all industry sectors and much work has been focused in developing novel techniques for big data analytics. Although many organisations see the potential of big data analytics, they are still in the initial stages of reaping its benefits since they are required to remodel their business processes and infrastructures to cater to the massive and fast-growing volume of big data [13]. Hence, research investigations were predominantly concentrating on cloud computing environment which can support big data storage and analytics due to its cost-effective scalability, elasticity, resource pooling and data processing

features [14–16]. However, cloud computing poses security and privacy risks [17,18].

Research studies were undertaken to provide data protection to maintain the integrity of data in the cloud and IoT environments [19,20]. However, big data systems are becoming more complex with distributed and heterogeneous environments calling for more research to address the growing security and privacy issues [21,22]. More recently, much research works in this direction have investigated in the specific context of health industry since big data is more prevalent in e-health care services [23,24]. However, with social and IoT networks, the big data security challenge has become an important part of the national agenda in many countries [25–27]. More and more personal data and user behaviours over the Internet are collected resulting in several privacy consequences that make existing privacy solutions inadequate for meeting different consumer contexts and privacy protection needs of users [28–30]. Further, research works have studied threat detection mechanisms in various networks, in particular social networks [30,31]. Recently, literature reviews were performed systematically on the main scientific peer reviewed journals on Scopus database, and big data papers were surveyed more specifically with a security and privacy perspective [8,32]. Results of such studies indicate that there is a strong focus on the computational aspects of big data security and privacy. It is important to note that such literature review studies serve only in identifying the prominent research categories and topics being studied, namely, privacy, data analytics and confidentiality. Only a handful of studies have provided an overview of big data and these have suggested new alternatives of research in the security and privacy area [12,21,33]. However, none of them have considered the security challenges of big data system holistically.

Overall, our literature review clearly indicates that there is a gap in research studies towards investigating the security and privacy concerns at various phases of big data life cycle by considering the big data system as a whole. Moreover, the evolution and integration of new disparate technologies introduces unprecedented challenges for big data privacy and security. There seems to be lack of research focus in exploring the fundamental problems of big data. It is important to understand the impending issues from first principles. Firstly, there is a need to unearth the various characteristics and dimensions of big data that are evolving with the recent technology innovations. Further, the evolving dimensions of big data exhibit certain influencing factors that can impact on big data security and privacy. These ramifications lead to new open issues in big data security that require a comprehensive study. The increasing security threats and challenges facing big data must be identified and addressed properly. To the best of our knowledge, previous research studies have not dealt with these issues using first principle's thinking. This forms the prime motivation and we aim to address the gap in literature with a modest first step in this paper.

3. The eleven Vs (11Vs) of big data

Historically, before the term big data was introduced to represent the vast amounts of data in the digitized world of today, the three commonly used dimensions characterising massive databases and data warehouses were the 3Vs, namely, Volume, Velocity, Variety [5]. However, with changes in technologies, big data has been characterised by additional dimensions that are more semantically applicable to big data per se. One of the foremost additional dimensions recognised is Veracity, which refers to the credibility or quality of data. Subsequently, Validity, Volatility and Value have formed the most relevant dimensions of big data [1,6]. Further, certain dimensions such as Variability

and Visualisation of big data have evolved, and more and more technical challenges are being identified in big data [34]. In particular, Valence and Vulnerability dimensions have been associated with the privacy and security challenges characterising big data of today. Due to the growing dimensions of big data beyond the traditional 3Vs, in this section, we describe the properties and characteristics of all the eleven dimensions to prepare ourselves for the security challenges surrounding the big data initiatives of the future. Figure 1 shows a summary of the 11Vs as dimensions identified in a typical big data system. These 11Vs are described below.

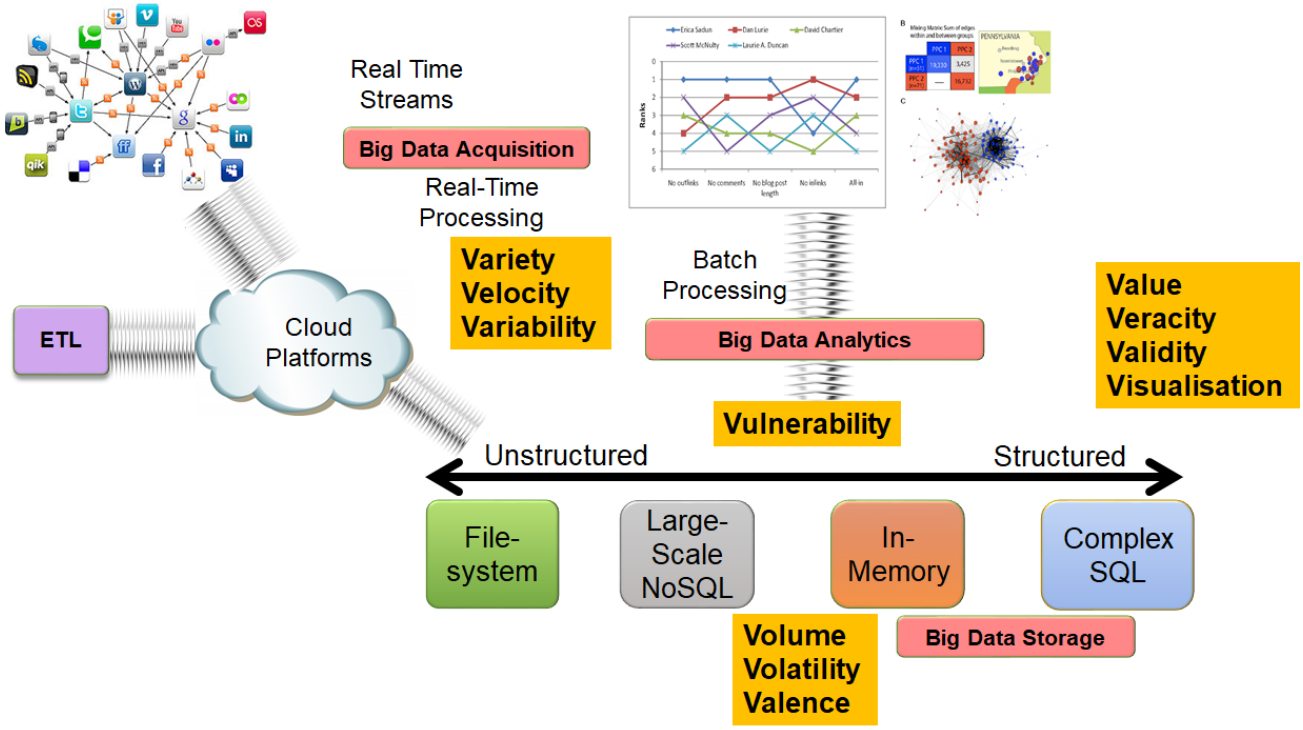

Figure 1. Dimensions of a typical big data system.

Dimension 1: Volume

The first known property of big data is Volume, which is attributed to the sheer size of data being collected [5]. With the developments of business websites getting integrated with social engineering and mobile applications (Apps), majority of today's data have been collected only recently, within the past couple of years [6,32]. There is an exponential growth in data generation and storage per day. Large multimedia files amounting to at least 300 hours of videos are being uploaded to YouTube and other social sites every minute [35]. Several social media sharing services are being leveraged by businesses in addition to their traditional transactional data, resulting in several trillions of data storage. The data volume of a single data source could be growing from petabytes to exabytes and zettabytes. A recent report by International Data Corporation (IDC) predicts that data would grow worldwide by 61% reaching 175 zettabytes in 2025 [36]. The volume of big data impacts the security and privacy in at least two major aspects as listed below:

i) data is stored in multiple locations (servers, nodes, clusters, etc.) in a distributed manner where conventional database systems and software tools are unable to continuously monitor and enforce standardised security protocols;

ii) any failure of a cluster or node can affect data transactions and performance within the tolerance time limits and is prone to security vulnerabilities.

Dimension 2: Velocity

The second dimension, Velocity deals with the speed at which new data gets generated and flows into organisations, and the increasing pace at which it needs to be processed in real-time [5,37]. The impact is on big data analytics where the rate of data creation needs to be matched with the real-time processing speed and capability of computing systems. Even though the volume of data storage could be enhanced, it is more important to consider the velocity at which new data is generated. Even if data is available, unless it can be processed in real-time, business opportunities could be lost. For instance, if weather predictions get delayed due to a slower processing speed that could not match with the velocity of data received, then it affects the right decisions required to be made at the right time. The velocity of big data impacts the security and privacy as faster cryptographic algorithms are required to keep up with the pace for real-time transaction processing. In addition, security audits are required to keep track of historical data by passing through privacy policies that match with the high rate of data accumulation.

Dimension 3: Variety

Big data exhibits heterogeneity with three types of forms such as structured, semi-structured, unstructured, which can be associated with its Variety dimension [5,33]. Majority of the data are unstructured that include files representing audio, video image, and sensor signals, as well as logs coming from social media, satellites, networks, and other machines. The Variety dimension does not refer to such different data representations alone, but also refers to the means and modes in which the same information is conveyed. While most common variety indicates the structural variety of data representation, it is also important to identify media variety or the different medium in which the same data is represented, and semantic variety indicating different meanings based on the different contexts of data. With structured data, a standard query using structured query language (SQL) could be employed to convey the associated semantic meaning, while unstructured data does not involve latent meaning. Recent adoption of email, XML, and other mark-up languages have led to a variety of semi-structured data. The variety of big data impacts the security and privacy with a need to have appropriate data classification and access controls for different data sources, types and formats.

Dimension 4: Veracity

Apart from the 3Vs characterising big data, with more and more uncertainty pertaining to data streaming and data availability, the credibility or quality of data has been considered to describe Veracity as the fourth important dimension of big data. For big data to be operational with a meaningful analysis, it is important to have the right and accurate data that can be processed in the right amount and at the right time [8]. Any data that is redundant incomplete or having errors cannot lead to good results when used in data analysis. When the first 3Vs increase with big data, the veracity reduces leading to less confidence or trust in the data. By improving the veracity of big data, the business risks associated with decision-making could be controlled. This has an impact on the security and privacy policies with respect to enforcing high quality data through appropriate application of data ownership and periodic access review methods.

Dimension 5: Validity

Another dimension related to Veracity is identified as the fifth dimension called Validity, which refers to the applicability of data with a specific context or intended use of the data. Hence, Validity establishes the correctness of the data for a specific use or view of the data in order to reap the benefits of big data analytics in contextual situations [1]. Since many organisations spend much time in cleaning the data before any data analysis can be performed, good data governance practices are required to maintain its validity as a continuous quality check process. This requires proper management of third-party vendors and partners enforcing protection of the entire data supply chain.

Dimension 6: Volatility

While Veracity and Validity dimensions characterise the quality assurance of big data, another dimension related to temporal aspects of the data is called Volatility, which determines how long the data is valid for it to be maintained in the data storage. This dimension ensures the currency of the datasets relevant for conducting real-time analysis [6]. Due to the cloud storage limitations and expenses associated with maintaining big data, robust policies for backup, and archiving are required to determine how long the data is to be held valid. In order to improve the performance of big data analytics, historic and irrelevant data should be archived regularly. The volatility of big data impacts the security and privacy policies and procedures for data retention, destruction and periodic re-assessment of security solutions.

Dimension 7: Value

The seventh dimension of big data is identified for understanding the benefits of big data associated with different stakeholders of an organisation that add Value to their business. This Value dimension refers to various factors that answers questions such as: which business decisions could leverage on big data insights, when is it most appropriate to make decisions, and who benefits from it directly [34]. In a nutshell, the Value dimension refers to measuring the usefulness of big data in making decisions for improving business performance. It helps organisations to embark on the right big data strategy so that big data analytics could help them to gain more data insights that are required for solving their complex business problems. Since analytics lead to action in businesses, the value of big data is an important dimension and appropriate access controls and approvals over analytical assessments are required. Also, determining appropriate security checkpoints during the development of such data insights are essential.

Dimension 8: Variability

All the above seven Vs could be affected by the eighth dimension of big data, namely Variability, which refers to inconsistencies in which variable data sources could load data into the data storage in variable speeds, formats or types [6]. It can also refer to outliers or anomaly detection that can benefit the organisation. When such information about the variability of big data is captured and associated with the data in the data storage, it can be utilised to make meaningful insights from big data. The IT security operations should cater to the big data variability aspects within the various audit log collections and monitoring methods.

Dimension 9: Visualisation

With different ways of data representation such as dashboards, heat maps, cone trees, and

k-means clustering, an important dimension of big data that has evolved more recently is its Visualisation for improving data insights [38]. Visualisation makes data easy to understand, and simple charts and graphs were traditionally used to communicate the data in a commonly acceptable graphical representation. Today, many sophisticated visualisation tools are being integrated with models for data analytics to make more meaningful graphical interpretations of big data in order to facilitate effective decision making. Hence, we consider Visualisation to be a popular requirement forming the ninth dimension of big data. Privacy and protection policies pertaining to the outputs from various visualisation tools should be established in addition to assigning access controls and privileges based on user roles and responsibilities.

Dimension 10: Valence

While big data could be massive, if the connections between the data items are not established, we would have pockets or islands of disparate data whose interrelationships may not be fully understood nor utilised. Any direct connections could be established when data gets collected as they get streamed. However, discovering indirect connections between data items is more difficult and they add value to the organisation. These interconnections, similar to the bonding between atoms in a molecule, result in the tenth dimension, namely Valence of big data [39]. It is a measure indicating how dense the data is, and a measure of Valence is determined as the ratio between the actually connected data items and the number of connections that could possibly be established within the data collection. Due to the large heterogeneous access points to data, scalability of hardware, network and systems is essential to maintain the appropriate valence that supports the service level agreement (SLA) of the big data system. Security and privacy management procedures should maintain the level of performance for both current and future growth of big data systems.

Dimension 11: Vulnerability

The last but the most important dimension is the Vulnerability property of big data, which relates to the security, privacy, and technology risks arising due to various rich personalised data collected through products and services using Internet applications, social networks, and IoT devices. The vulnerabilities due to security and privacy gaps and the absence of standards are associated to big data technologies, processes and management [40]. Recently reported breaches of security and privacy of big data has drawn much attention to look into the Vulnerability dimension of big data [30,41]. Security and privacy policies and procedures for incident management with regard to big data are driven by the abovementioned dimensions that require continuous monitoring. Periodic vulnerability checks and penetration tests are to be developed catering to the unique features of big data. The vulnerabilities of sensitive data leakage must be identified and appropriate measures to review the confidentiality, integrity and availability of big data systems and data are required.

4. Security and privacy challenges of big data

In this big data era, cloud storage services are being used extensively to cope with the exponential growth of data [2,14]. More data is accessible, and organisations gain much knowledge with big data mining that can aid in their intelligent business decision making [6,13]. While big data can bring many value-added opportunities to businesses and people in their daily operations, it also creates security and privacy challenges due to its use in intelligent services. Novel and easy-to-use

products and services such as the Internet, IoT and smart devices, as well as social networks have also aroused the attackers to take this platform as an opportunity to impose malicious activities for financial benefits [20,30]. Rich information about a person's interest, preference, movement, and behavior patterns could be captured, and such sensitive and valuable data could be leaked. Hence, appropriate protection and privacy need to be enforced throughout the lifecycle of big data. In this section, we analyse the security and privacy challenges of big data in each of the three phases of its lifecycle: i) big data acquisition, ii) big data storage, and iii) big data analytics. A summary of the security and privacy challenges of big data and their impact in these three phases are given in Figure 2, which is described below.

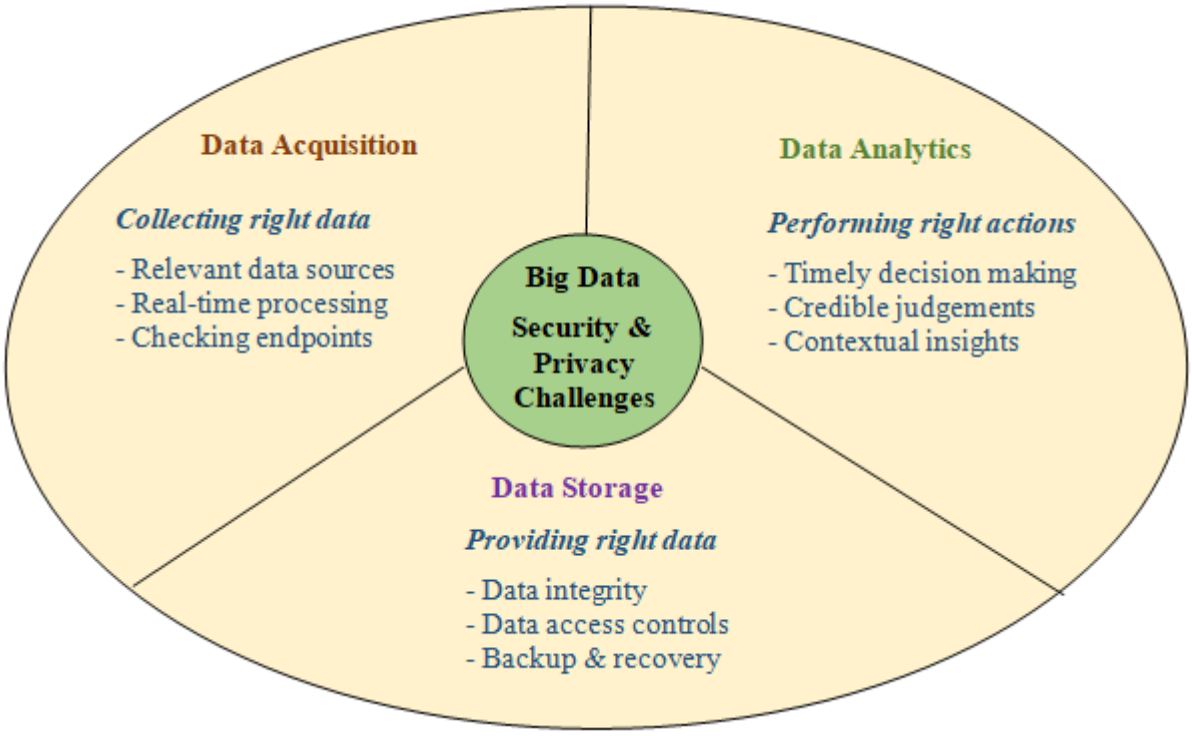

Figure 2. Security and privacy challenges of big data and their impact.

Big data acquisition

Big data is acquired from different sources and media via online services and business transactions, where some are structured, but many are unstructured and semi-structured including real-time processed data [27,33]. We have identified key challenges in this phase that are mapped to the prominent Vs of big data as follows:

- due to the Variety of data, traditional security techniques that adopt encryption methods are not suited for all types of sources from which big data acquisition takes place.
- the Velocity of acquisition of data is so high that it becomes difficult to monitor the traffic in real-time as data gets streamed into the storage [41].
- due to the Variability dimension of big data there can be inconsistencies in data formats, speed and types, such as in Web clickstream data leading to security vulnerabilities.

Overall, due to the above issues, big data when acquired could possibly become the carrier of advanced persistent threat (APT). APT thrives in such situations where there is a variety of data

sources and non-standard data formats for data streams coming as an ongoing process in social networks [30]. When an APT code is hidden in big data, it becomes difficult to be detected in real-time. Hackers could attack the data source, destination, and all the connectivity by capitalising on their vulnerabilities, which could result in an enlarged attack by launching a botnet. Therefore, it is important to enforce data security and privacy policies within a real-time processing environment of big data during the data acquisition stage itself. It is essential to connect the right endpoints of a network for the data flow along with sophisticated authentication and privacy policies for big data.

Big data storage

In this phase of big data life cycle, the data storage collects and maintains the data acquired from various data sources. The main purpose of having a big data storage is to help the internal and external stakeholders of the organisation in retrieving any relevant information in the future, whenever required. The key challenges of this phase are as follows:

- With the growing massiveness of big data, the Volume dimension affects the server infrastructure of an organisation. Traditional data warehouses may not suffice, and alternate storage systems such as distributed, cloud, and other outsourced big data servers are required to be employed to cope with the volume as well as the increasing Velocity of big data storage [19,20].
- Structured, unstructured, and semi-structured data are getting accumulated in the big data storage with high Volatility from various channels, including online transactions of sales, customer feedback, social media messages, emails, marketing information, and various other logs that are both directly and indirectly associated with the business operations of an organisation.
- Big data is also shared among multiple related departments for their day-to-day transactions and functional operations. Hence, the big data connectivity among different data centres that are in-house, cloud-based, or outsourced can make the big data quite dense, impacting on the Valence dimension of big data.

Due to the above, the integrity of the data could be affected when multiparty operations take place on the same data storage in huge amounts and in increasing real-time speed [16]. Traditional encryption and security measures to maintain data integrity may not help as multiple mechanisms could be in place. Such a disparate environment could encourage sniffers to reach the servers by exploiting their security policy differences and vulnerabilities. Any misuse of data could lead to privacy leakage. These factors increase the risk of information theft and user privacy infringement.

Traditional access control methods are mainly classified as mandatory, discretionary or role-based, and none of them can be effectively applied to big data storage due the diversity of users and their permission rights in such a highly dynamic environment. Hence, new trustworthy data access controls must be established, adhering to appropriate security and privacy protection schemes and policies [42]. Good practices for backup and recovery must be followed in dealing with historic data that require to be archived or destroyed at every stage of the big data life cycle.

Big data analytics

The purpose of taking the effort to collect and maintain big data with good data storage systems is to use, process and analyse big data for gaining data insights and to make timely as well as accurate decisions. So, the primary aim is to perform big data analytics, which is an important phase in the life cycle of big data. Organisations collect many contextual and sensitive information about customers to analyse their interactions in order to arrive at a more meaningful marketing strategy for

providing them with personalized products and enhanced services. Analysing such data in the cloud from a variety of sources could lead to unexpected privacy leakage [43]. There are typically three main steps in data analytics: i) data preparation to identify, clean and format the data according to the requirements of the analytics model; ii) adoption of the analytic model; and iii) communication of the output to provide data insights. Each of these steps face many security and privacy challenges due to their inherent vulnerabilities.

Patient Id	Age ▲	Gender	Flu Patient	Flu Disease Patient	Hypertension Patient	Stroke Patient	BMI	Primary Med
21954120094	50	M	YES	NO	NO	NO	38.43	DPP-4 Inhibitor
86208842988	50	F	YES	NO	NO	NO	36.79	Thiazolidinedione
42164940704	51	F	YES	NO	NO	NO	41.87	Thiazolidinedione
35585823647	51	M	YES	NO	YES	NO	45.21	Biguanide
50199370537	52	F	NO	NO	YES	NO	37.50	
50491773653	52	M	NO	NO	YES	NO	40.46	
74990562311	53	M	NO	NO	NO	NO	45.89	

Figure 3. Complex visualisation with big data insights.

We identify four key challenges in the big data analytics phase as follows:

i) the Value of the results obtained from big data analytics is based on how reliable the data is. If the data comes from untrusted source or has been tampered in the storage systems, data analytics would lead to wrong judgements [44].
ii) anonymising the data is not foolproof and user privacy can get revealed with big data analytics. The credibility of the data analytics is questionable when the Veracity dimension of big data is affected.
iii) big data analytics makes use of machine learning and other intelligent algorithms to process huge data for a specific context. This Validity dimension of big data could be affected by hackers. For instance, new applications are being developed to filter emails for identifying a topic or if the email is a spam. A dictionary of legitimate words is created in this process. Hackers take the opportunity to hide malicious code in such applications so that bogus analytics could reveal sensitive information and contextual information from big data.
iv) there could be distortion in communicating the results of big data analytics, especially with a hampered Visualisation dimension of big data analytics. For instance, Figure 3 provides an illustration of a complex visualisation tool depicting rich data about influenza flu demographics with lots of buried sensitive information and decision parameters that can provide the necessary big data insights for decision makers in a hospital environment. Such data visualisation could be available in just a click away to anyone who has access to the visual tools or the artefacts, if appropriate security and privacy policies are not enforced.

Overall, hackers could pose a threat to any of the steps of big data analytics to target an attack. Several studies indicate that hackers obfuscate malicious code to evade detection and such evasion attacks affect the data integrity and result in indiscriminate violation of user privacy. Initially, such attacks were not targeting machine learning algorithms. However, more recent research studies have shown that attackers inject poison in training schemes of a feature-based machine learning algorithm to affect feature selection or inject a classification error in order to deviate a user's topic towards the attacker's preference [45]. When data insights are not clearly and accurately presented using appropriate visualisation tools, businesses will not be able to perform the right action. Business reports and dashboards bury nuggets of information using complex graphics and visualisation techniques so that it becomes unrealistic to detect any hidden attacks in such data presentations. Some of the common challenges of visualisation are information loss, steganography, visual noise and fast image changes [38]. Visual privacy is also becoming an interesting research topic.

In summary, an attacker's goal is to exploit the vulnerabilities present in all the above mentioned three phases of big data life cycle to violate the integrity, availability, and privacy of data, either with a specific target or an indiscriminate target. This can have a spectrum of financial, personal and psychological impact on individuals and organisations.

5. Proposed big data security strategies

New storage and computing infrastructures such as cloud platforms and NoSQL databases that are introduced to cope with the 11Vs of big data are also becoming the vehicles for new attacks. In the previous section, we have identified the key security and privacy challenges that are specific to big data during the three phases of big data life cycle. In this section we propose big data security

strategies by making use of the following popular and contemporary technologies that can be adapted to the new paradigm of big data. We propose the use of four main technologies to comprehensively cope with the 11Vs during the three phases of big data lifecycle, namely data acquisition, data storage and data analytics. These are i) data provenance technology, ii) data encryption and access control technology, iii) data mining technology, and iv) blockchain technology.

i) Data provenance technology – to adapt data provenance technology for addressing the security and privacy challenges in the data acquisition phase of big data.

In traditional computing systems, data provenance method was used to determine the source of the data in data warehouse by adopting the labelled technique. With big data, the data acquisition involves diverse data sources from the Internet, cloud, social, and IoT networks. While big sensing data streams come with novel encryption schemes, attacks are possible right from the data acquisition phase [46]. Hence, metadata about these data sources such as the data origin, the process used for dissemination, and any intermediate calculations could be recorded in order to facilitate mining of the information at the time of data streaming itself. Hence, the first strategy we propose is to adapt data provenance technology for effectively using data analytics techniques for detecting anomalies in the data acquisition phase of big data. However, collecting provenance metadata must adhere to privacy compliance. Another important issue is that it could become complex with application tools generating growingly large provenance graphs for establishing metadata dependencies. Data analytics of such graphs could be computationally intensive, and algorithms are being developed to detect anomalies using proximity graphs [31,47]. For instance, within a data provenance technology, the social network analysis component could adopt a machine learning anomaly detection model based on a social score using the following graph metrics:

- Degree Centrality (DC) = number of nodes are connected to v

$$DC(v) = deg(v)$$

- Betweenness Centrality (BC) = number of paths, from a (source, destination) or (s, t) pair that go through v

$$BC(v) = \sum_{s,t \in V} \left(\frac{p(s,t/v)}{p(s,t)}\right)$$

Apart from balancing the trade-offs between privacy and computational complexity, it is important to monitor the data provenance technology itself as it could be attacked and requires security protection from hackers. For instance, in a real-life example of identifying the owner of the provenance documents, provenance graphs with chains of derivations of the vocabulary that contain the provenance information could have variations based on the provenance producer's individual style. The above-mentioned graph metrics could be used in provenance network analytics to identify the producer of the provenance graph [48].

ii) Data encryption and access control technology – to adapt advanced encryption techniques and access control schemes in big data storage systems.

Contemporary schemes such as homomorphic, attribute-based, and image encryption are being explored to ensure that sensitive private data is secured in cloud and other big data storage and service platforms [49,50]. Even though homomorphic encryption allows some operations on

encrypted data without decrypting it, the computing efficiency and scalability of homomorphic encryption schemes need improvement in order to be able to handle big data. On the other hand, the attribute-based encryption technique is regarded more appropriate for end-to-end security of private data in the cloud environment since the decryption of the encrypted data is possible only if a set of attributes of the user's secret key matches with the attributes of the encrypted data [50]. One of the major challenges of this scheme is the implementation of revocation since each attribute may belong to different multiple set of users [11,51]. Anonymising data with a hidden key field could be useful for privacy protection. However, using data analytics such as correlation of data from multiple sources, an attacker would be able to identify the anonymised data. Hence, in addition to having good cryptographic techniques to ensure privacy and integrity of active big data storage, proof of data storage needs to be continuously ensured. Another important aspect is to provide proof of the archived data storage in order to verify that the files are not deleted or modified by attackers. Some methods for such granular audits use indirect homomorphic RSA-based hash introduced by Shamir to ensure the proof of data storage or to demonstrate that an arbitrary set of data possessed by a person is known to a verifier [52]. The mathematical model can be expressed as follows:

Let N be an RSA modulus with

$N = pq$, where p and q are large prime numbers.

Suppose F is the dataset (or file) to be verified and is represented as an integer for the mathematical modeling of the method. The verifier stores k for the dataset (or file) F, and is given by

$$k = F \bmod \varphi(N)$$

The verifier sends a random element $g \in Z_N$, and the person in possession of the dataset returns

$$s = g^F \bmod N$$

Next, the verifier needs to check if $g^k \bmod N = s$ in order to confirm authentication and proof of recoverability of datasets from archived data storage. Variations of the scheme with efficient computations using data block positions and attribute-based access control schemes with dynamic policy updates are being researched for big data in the cloud environment For instance, let us consider a real-life example of enforcing the above data encryption and access control techniques with health data in IoT devices. In such a context, it is demonstrated that a combination of attribute-based encryption schemes with outsourced key generation in the cloud could be adopted in order to overcome the limitations of data storage and RSA-based computation capability in IoT devices [53].

iii) Data mining technology – to adapt data mining techniques within big data analytics to intelligently perform behaviour mining of access controls, authentication and incident logs.

Data mining technologies are on the rise to identify vulnerabilities and risks in big data and to predict any threats as a prevention technique from any possible malicious attack [13,45,54]. Role mining algorithms automatically extract and optimise the roles that can be automatically generated based on the user's access records for efficiently providing personalized data services for mass users. However, in big data environment, it is important to ensure the dynamic changes and the quality of data pertaining to the roles assigned to users and roles related to the permissions-set that simplify rights management. In big data environment, it may not be possible to accurately specify the data which users can access. In such a context, adopting risk-adaptive access controls using statistical

methods and information theory would be applicable. However, defining and quantifying the risks are quite difficult. Hence, authentication based on behaviour characteristics of users could be adopted, but the big data system needs to be trained with the training dataset as a continuous process. Incident logs pertaining to the Intranets, Internet, social, and IoT networks as well as email servers could be analysed to detect abnormal behaviour or anomaly patterns using appropriate data mining techniques [30,31]. While traditional threat analysis cannot cope with big data, by using behaviour mining of metadata of various resource pools related to big data, anomalies can be analysed to predict the threats, such as an APT attack.

In behaviour mining, trend analysis is performed, and pattern proximity is measured to define the relation between datasets. A distance function is usually used to measure the pattern proximity [47]. The distance function defines the proximity between two datasets based on their attributes. It is obvious that a group of datasets which has the minimum distance value between them belong to the same cluster. The most popular general distance function $d_{i,j}$ between two datasets x_i and x_j with p attributes is the Minkowski distance metric in the normed vector space of order m, and is used to calculate the pattern proximity as follows:

$$d_{ij} = \sqrt[m]{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^m}$$

When $m = 2$, the Minkowski distance is the commonly used Euclidean distance metric as follows:

$$dist(\boldsymbol{x_i}, \boldsymbol{x_j}) = \|\boldsymbol{x_i} - \boldsymbol{x_j}\| = \sqrt{(\boldsymbol{x_i} - \boldsymbol{x_j}).(\boldsymbol{x_i} - \boldsymbol{x_j})} = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2}$$

The Euclidean function works well when the datasets exhibit compact or isolated clusters and is suitable for patterns with multiple dimensions. Big data security can be enhanced by studying the pattern proximity to predict threats by training with the similarity metrics of distances between anomaly datasets and normal datasets based on server/network logs, historical data of incidents and social media data. However, threat detection schemes require scalability and interoperability for the big data environment.

The parallel framework offered by MapReduce for proximity mining is a good fit for implementing data mining technologies as it can perform efficient data-intensive computations and machine learning in providing high scalability and support with large heterogeneous data sources. MapReduce uses two functions called Map and Reduce that process list of pairs <key, value>. The Map function inputs a list of keys and associated values and produces a list of intermediate <key, value> pairs. As shown in Figure 4, the feature attributes determined to perform proximity mining between datasets are used for representing the <key, value> pairs in the MapReduce framework. Next, grouping and shuffling of intermediate pairs are performed in proximity mining for determining Matched Pairs and Matched Classes by adapting the distance measures for similarity metrics. Finally, the Reduce function performs the merge operation for the Fusion of Classes on all intermediate pairs to arrive at the final result. The MapReduce framework can be implemented using a Hadoop Distributed File System (HDFS) as it provides parallel and scalable architecture for a large-scale data storage based on clusters in the cloud [35]. However, Hadoop systems are also being attacked as data leakages in the data mappers could take place either intentionally or unintentionally in a cloud cluster.

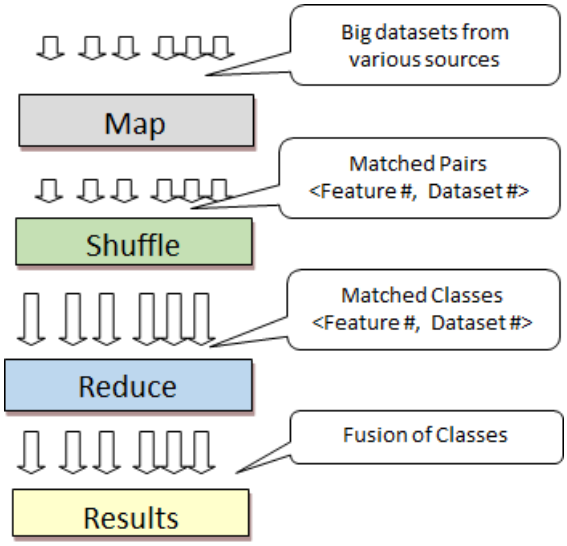

Figure 4. MapReduce framework for proximity mining of big datasets.

For instance, in real-life scenarios, a common method to protect the servers from malicious attacks is to record server logs that are analysed for anomalies. In such contexts, a MapReduce model for distance-based anomaly analysis can be deployed to find k nearest neighbors for a data point and to use its total distance to the k nearest neighbors as the anomaly score in a data mining algorithm [55]. In order to perform the anomaly detection task, the MapReduce functionality can be divided into two jobs as given below:

i. a MapReduce job to find pair wise distance between all data points.
ii. a MapReduce job to find the k nearest neighbors of a data object and to find the weight of the object with respect to the k nearest neighbors.

Data objects can be partitioned by hashing the object ID and all possible combination of hash values for every two object types can be considered. The Euclidean distance computation between objects for each hash value pair can be distributed among the reducers with the mapper output key as a function of the two hash values. However, configurable number of hash buckets should be chosen appropriately to distribute the load uniformly across the reducers by using Hadoop's default reducer partitioner. The parallel processing feature of Hadoop speeds up the processing time resulting in a real-time data mining of large datasets of log files efficiently for detecting anomalous events in the server.

iv) Blockchain technology – to adapt distributed trusted system based on blockchain for big data security and privacy protection.

Blockchain technology has demonstrated to be the new mode of trusted interaction and exchange of information by eliminating intermediate parties and supporting direct communication between two parties in a network through replication of information and validation processes [56]. A blockchain is a shared ledger technology ensuring appropriate visibility for any participant in the business network to see the system of record (ledger). In a blockchain system, all transactions are secure, authenticated and verifiable as all parties agree to a network verified transaction and is suitable for applications that need trust with properties such as consensus, immutability and provenance [57,58]. Overall, it can be well-suited for big data security for various organisations.

A blockchain represents a decentralized peer-to-peer network with a historical archive of decisions and actions undertaken. Blocks retain data via a list of transactions and are chained together through each block containing the hash of the previous block's header, thus forming the blockchain with inherent encryption process. Figure 5 shows a generalised blockchain in operation with blocks storing the proof of transactions, timestamp, and other information on a user activity. Blockchains are maintained through the consensus of a set of nodes running blockchain software, called mining nodes.

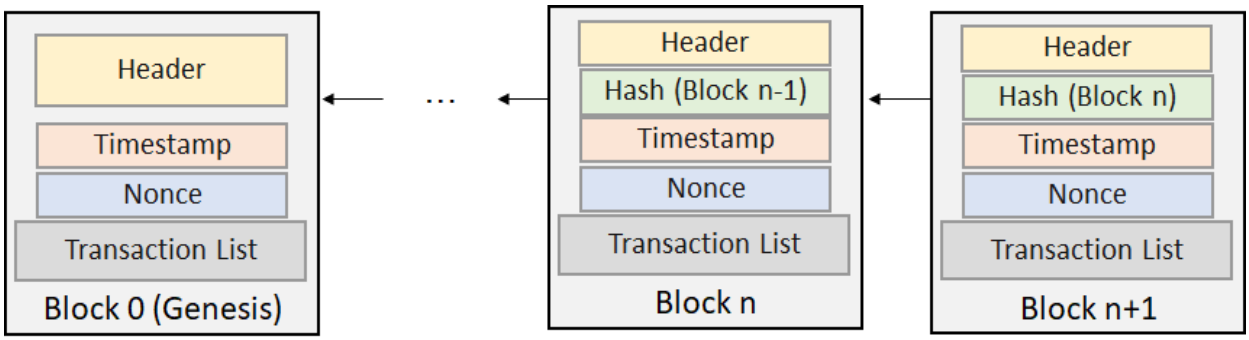

Figure 5. Generic blockchain in operation.

In big data context, each data item or record in a database is a block containing transaction details including the transaction date and a link to the previous block. The integrity of data is maintained in blockchain technology. This is because corrupted data cannot enter into the blockchain as checks are carried out continuously in search of patterns in real-time by the various computers on the network. Also, blockchain allows sharing of data more wisely as contracted by the users, thereby preventing cybercrime and data leakage. Blockchain data could also provide valuable insights into the behaviours, trends and can be used for predictive analytics.

In a typical business environment, there is partial trust within a company or between companies, and IBM's Hyperledger Fabric blockchain could be adopted. A recent study used such a blockchain technology to implement access control of big data by combining two existing access control paradigms: 1. Identity-Based Access Control (IBAC), and 2. Role-Based Access Control (RBAC) [59]. The Blockchain Identity-Based Access Control Business Network (BIBAC-BN) uses the Hyperledger composer that consists of a model resource with definitions of a person participant, data asset with access control of five operations implemented: 1. Request access, 2. Grant access, 3. Revoke access, 4. Verify access, and 5. View asset. For the BRBAC-BN model, the resource file contains definitions of the persons and all organisation participants so that access privileges are granted to a subset of the big data based on their roles. This way, users' roles enable them to access an asset only if it is enabled on the blockchain for the assigned specific roles.

Overall, the 11Vs of big data encompassing the three phases of its life cycle, namely big data acquisition, big data storage, and big data analytics can have the additional layer of security when implemented on a blockchain network. However, we need to also consider some of the concerns associated with the blockchain technology as listed below:

- Irreversibility – encrypted data may be unrecoverable when the private key of a user is lost.
- Adaptability challenges – organisations need to adapt the technology for integrating it in their existing supply chain systems, which may require a big change management and learning curve.

- Current limitations – there are high operational costs associated with running blockchain technology as it requires expert developers, substantial computing power, and revamping resources to cater to its storage limitations
- Risks and Threats – while the blockchain technology greatly addresses the security challenges of big data, it is not threat proof. If the attackers are able to penetrate into majority of the network, then there is a risk of losing the entire database.

6. Concluding remarks

The dynamic integration of different technologies via Intranets, cloud infrastructures, the Internet, social, and IoT networks has resulted in highly complex and heterogeneous big data systems. This paper attempted to satisfy the need to underpin the evolving new privacy and security issues using a holistic approach of applying first principles thinking for understanding the entire big data system. Firstly, we established the premise of big data environment in order to understand its complexities for addressing these challenges. From the evolution of various dimensions of big data, we identified eleven key dimensions or 11Vs namely, Volume, Velocity, Variety, Veracity, Validity, Volatility, Value, Variability, Visualisation, Valence and Vulnerability, which constitute the main characteristics that have a direct or indirect impact on the escalating privacy and security issues in big data. Secondly, we considered the big data system life cycle with three main phases, namely big data acquisition, big data storage and big data analytics. Further, the 11Vs of big data were mapped to each phase based on the privacy and security perspectives. Finally, we provided four practical strategies spanning every phase of big data life cycle by using the application of certain contemporary technologies to address the privacy and security challenges of big data. Four existing technologies using popular techniques of data provenance, data encryption and access control, data mining and blockchain were discussed with suitable adaptation in order to address the security challenges encountered throughout the big data life cycle. Overall, this paper unearthed a number of challenges, open issues and possible technology adoption strategies in the big data security arena. The underlying ramifications could trigger future research in this important topic.

Acknowledgments

The authors wish to thank the affiliated educational institutions for the encouragement and support given to this research work.

Conflict of interest

The authors declare no conflict of interests.

References

1. M. Chen, S. Mao and Y. Liu, *Big Data: A Survey*, Mobile Netw. Appl., **19** (2014), 171–209.
2. W. Tian and Y. Zhao, *Big data technologies and cloud computing*, Optimized Cloud Resource Management and Scheduling Theory and Practice, (2015), 17–49.

3. C. L. McNeely, J. Hahm, *The big (data) bang: policy, prospects, and challenges*, Review of Policy Research, **31** (2014), 304–310.
4. A. Gandomi, M. Haider, *Beyond the hype: Big data concepts, methods, and analytics*, International Journal of Information Management, **35** (2015), 137–144.
5. D. Laney, *3D Data Management: Controlling Data Volume Velocity and Variety*, META Group research note, **6** (2001), 1.
6. J. Frizzo-Barker, P. A. Chow-White, M. Mozafari, et al. *An empirical study of the rise of big data in business scholarship*, International Journal of Information Management, **36** (2016), 403–413.
7. T. Huang, L. Lan, X. Fang, et al. *Promises and challenges of big data computing in health sciences*, Big Data Res., **2** (2015), 2–11.
8. B. Nelson, T. Olovsson, *Security and privacy for big data: A systematic literature review*. In: 2016 IEEE International Conference on Big Data (Big Data), (2016), 3693–3702
9. M. Li-chuan, P. Qing-qi, L. Hao, et al. *Survey of Security Issues in Big Data*, Radio Communications Technology, **41** (2015), 1–7.
10. F. Deng-Guo, Z. Min, L. Hao, *Big Data Security and Privacy Protection*, Chinese Journal of Computers, **37** (2014), 246–258.
11. N. B. Kshetri, *The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns*, Big Data & Society, **1** (2014), 1–20.
12. X. Jin, B. Wah, X. Cheng, et al. *Significance and challenges of big data research*, Big Data Research, **2** (2015), 59–64.
13. W. Xindong, Z. Xingquan, W. Gong-Qing, et al. *Data Mining with Big Data*, IEEE T. Knowl. Data En., **26** (2014), 97–107.
14. V. Chang and G. Wills, *A model to compare cloud and non-cloud storage of Big Data*, Future Gener. Comp. Sy., **57** (2016), 56–76.
15. Z. Goli-Malekabadi, M. Sargolzaei-Javan, M. K. Akbari, *An effective model for store and retrieve big health data in cloud computing*, Comput. Meth. Prog. Bio., **132** (2016), 75–82.
16. N. Kumar, A. V. Vasilakos, and J. Rodrigues, *A multi-tenant cloud-based DC nano grid for self-sustained smart buildings in smart cities*, IEEE Commun. Mag., **55** (2017), 14–21.
17. S. Subashini and V. Kavitha, *A survey on security issues in service delivery models of cloud computing*, J. Netw. Comput. Appl., **34** (2011), 1–11.
18. H. Cheng, W. Wang, and C. Rong, *Privacy protection beyond encryption for cloud big data*. In: Proceedings of the 2nd International Conference on Information Technology and Electronic Commerce, (2014), 188–191, IEEE.
19. P. Jing, *A new model of data protection on cloud storage*, Journal of Networks, **9** (2014), 666–671.
20. C. Liu, C. Yang, X. Zhang, et al. *External integrity verification for outsourced big data in cloud and IoT: a big picture*, Future Gener. Comp. Sy., **49** (2015), 58–67.
21. H. Kun, L. Di, L. Minghui, *Research on Security Connotation and Response Strategies for Big Data*, Telecommunications Science, **30** (2014), 112–117.
22. T. Matzner, *Why privacy is not enough privacy in the context of ubiquitous computing and big data*, Journal of Information, Communication and Ethics in Society, **12** (2014), 93–106.
23. D. Thilakanathan, Y. Zhao, S. Chen, et al. *Protecting and Analysing Health Care Data on Cloud*. In: Proceedings of the 2nd International Conference on Advanced Cloud and Big Data, (2014), 143–149, IEEE.

24. I. de la Torre-D éz, B. Garcia-Zapirain, M. Lopez-Coronado, et al. *Proposing telecardiology services on cloud for different medical institutions: a model of reference*, Telemedicine and e-Health, **23** (2017), 654–661.
25. G. Lafuente. *The big data security challenge*, Network Security, **2015** (2015), 12–14.
26. R. Lu, H. Zhu, X. Liu, et al. *Toward efficient and privacy-preserving computing in big data era*, Network IEEE, **28** (2014), 46–50.
27. J. W. Crampton, *Collect it all: national security, Big Data and governance*, GeoJournal, **80** (2015), 519–531.
28. D. Lyon, *Surveillance, snowden, and big data: Capacities, consequences, critique*, Big Data & Society, **1** (2014), 1–13.
29. X. Hu, M. Yuan, J. Yao, et al. *Differential Privacy in Telco Big Data Platform*, Proceedings of the VLDB Endowment, **8** (2015), 1692–1703.
30. M. Benjamin, S. B. Michelle and T. B. Nadya, *Eigenspace Analysis for Threat Detection in Social Networks*. In: 14th International Conference on Information Fusion, (2011), 1–7, IEEE.
31. A. Leman, T. Hanghang, K. Danai, *Graph based anomaly detection and description: a survey*, Data Min. Knowl. Disc., **29** (2015), 626–688.
32. C. Rebello, E. Tavares, *Big Data Privacy Context: Literature Effects On Secure Informational Assets*, Transactions on Data Privacy, **11** (2018), 199–217.
33. C. R. Silva, E. M. T. Rodrigues, *Privacy In Big Data: Overview And Research Agenda*, Sistemas & Gestao, **12** (2017), 491–505.
34. M. A. Khan, M. F. Uddin, N. Gupta, *Seven V's of Big Data Understanding Big Data to extract Value*. In: Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education, (2014), 1–5.
35. S. Hota, *Big Data Analysis on YouTube Using Hadoop And Mapreduce*, International Journal of Computer Engineering In Research Trends, **5** (2018), 98–104.
36. A. Patrizio, *IDC: Expect 175 zettabytes of data worldwide by 2025*, Network World, 2018.
37. K. D. Gronwald, *Big Data Analytics*. In: Integrated Business Information Systems A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data, (2017), 127–157.
38. E. Y. Gorodov and V. V. Gubarev, *Analytical review of data visualization methods in application to big data*, Journal of Electrical and Computer Engineering, **2013** (2013), 22.
39. G. B. Tarekegn, Y. Y. Munaye, *Big Data: Security Issues, Challenges and Future Scope*, International Journal of Computer Engineering & Technology, **7** (2016), 12–24.
40. F. Almeida, *Big Data: Concept, Potentialities and Vulnerabilities*, Emerging Science Journal, **2** (2018), 1–10.
41. B. A. Kumar, S. Maninder, *Data mining-based integrated network traffic visualization framework for threat detection*, Neural Computing and Applications, **26** (2015), 117–130.
42. Z. Yan, W. Ding, V. Niemi, et al. *Two schemes of privacy-preserving trust evaluation*, Future Gener. Comput. Sy., **62** (2016), 175–189.
43. N. Rastogi, M. J. K. Gloria, J. Hendler, *Security and Privacy of Performing Data Analytics in the Cloud*, Journal of Information Policy, **5** (2015), 129–154.
44. E. Bozdag, *Bias in algorithmic filtering and personalization*, Ethics and information technology, **15** (2013), 209–227.
45. H. Xiao, B. Biggio, G. Brown, et al. *Is Feature Selection Secure against Training Data Poisoning?* International Conference on Machine Learning, (2015), 1689–1698.

46. D. Puthal, S. Nepal, R. Ranjan, et al. *A dynamic prime number based efficient security mechanism for big sensing data streams*, J. Comput. Syst. Sci., **83** (2017), 22–42.
47. Y. Zhe, M. Philip and R. Michael, *Anomaly Detection Using Proximity Graph and PageRank Algorithm*, IEEE T. Inf. Foren. Sec., **7** (2012), 1288–1300.
48. T. D. Huynh, M. Ebden, J. Fischer, et al. *Provenance Network Analytics: An approach to data analytics using data provenance*, Data Min. Knowl. Disc., **32** (2018), 708–735.
49. G. Zhou, D. Zhang, Y. Liu, et al. *A novel image encryption algorithm based on chaos and line map*, Neurocomputing, **169** (2015), 150–157.
50. Z. Wang, C. Cao, N. Yang, et al. *ABE with improved auxiliary input for big data security*, J. Comput. Syst. Sci., **89** (2017), 41–50.
51. C. Hsu, B. Zeng and M. Zhang, *A novel group key transfer for big data security*, Appl. Math. Comput., **249** (2014), 436–443.
52. D. L. G. Filho and P. S. L. M. Barreto, *Demonstrating data possession and uncheatable data transfer*, IACR Cryptology ePrint Archive, **2006** (2006), 150.
53. K. P. Kibiwott, Y. Zhao, J. Kogo, et al. *Verifiable fully outsourced attribute-based signcryption system for IoT eHealth big data in cloud computing*, Mathematical Biosciences and Engineering, **16** (2019), 3561–3594.
54. G. Fuchs, H. Stange, D. Hecker, et al. *Constructing semantic interpretation of routine and anomalous mobility behaviors from big data*, SIGSPATIAL Special, **7** (2015), 27–34.
55. G. Remya, A. Mohan, *Distributed Computing Based Methods for Anomaly Analysis in Large Datasets*, International Journal of Advanced Research in Computer and Communication Engineering, **4** (2015), 427–430.
56. F. Restuccia, S. D. Kanhere, T. Melodia, et al. *Blockchain for the Internet of Things: Present and Future*, IEEE Internet of Things Journal, **1** (2018), 1–8.
57. K. Christidis, M. Devetsiokiotis, *Blockchains and Smart Contracts for the IoT*, IEEE Access, **4** (2016), 2292–2303.
58. D. Yaga, P. Mell, N. Roby, et al. *Blockchain Technology Overview*, National Institute of Standards and Technology, U.S. Department of Commerce, (2018), 1–27.
59. U. U. Uchibeke, K. A. Schneider, S. H. Kassani, et al. *Blockchain Access Control Ecosystem for Big Data Security*. In: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), (2018), 1373–1378.

AIMS Press