**Noname manuscript No.**
(will be inserted by the editor)

# Towards Big Data Governance in Cybersecurity

**Longzhi Yang · Jie Li · Noe Elisa · Tom Prickett · Fei Chao**

Received: date / Accepted: date

**Abstract** Big data refers to large complex structured or unstructured data sets. Big data technologies enable organisations to generate, collect, manage, analyse, and visualise big data sets, and provide insights to inform diagnosis, prediction, or other decision-making tasks. One of the critical concerns in handling big data is the adoption of appropriate big data governance frameworks to: 1) curate big data in a required manner to support quality data access for effective machine learning, and 2) ensure the framework regulates the storage and processing of the data from providers and users in a trustworthy way within the related regulatory frameworks (both legally and ethically). This paper proposes a framework of big data governance that guides organisations to make better data-informed business decisions within the related regularity framework, with close attention paid to data security, privacy and accessibility. In order to demonstrate this process, the work also presents an example implementation of the framework based on the case study of big data governance in cybersecurity. This framework has the potential to guide the management of big data in different organisations for information sharing and cooperative decision-making.

This work has been supported by the Royal Academy of Engineering (IAPP1\100077), and the Commonwealth Scholarship Commission (CSC-TZCS-2017-717).

Longzhi Yang, Jie Li, Noe Elisa and Tom Prickett
Department of Computer and Information Sciences
Faculty of Engineering and Environment
The University of Northumbria at Newcastle
E-mail: longzhi.yang@northumbria.ac.uk

Fei Chao
Cognitive Science Department
Xiamen University
E-mail: fchao@xmu.edu.cn

**Keywords** Big data governance · Cybersecurity · Data quality management · Artificial intelligence · Big data analysis

## 1 Introduction

The growth in interconnected networks and devices has resulted in an explosive data increase in organisations. The data is increasingly used to provide insights by analytics, which inform critical business decisions. The ongoing digitisation of commercial and non-commercial organisations has contributed to this growth, as has the increasingly wide use of Internet of Things (IoT). IoT devices gather information from various sectors, such as health, energy, weather, business, transportation, education and manufacturing [1], and intend to make positive impacts to the society and the environment. The large amount of information is commonly referred to as 'big data', that is collected, mined, analysed, and visualised in order to find behavioural trends and patterns to inform decision-making [2].

The common challenges associated with big data are to store and analyse the collected datasets, to provide insights in a timely manner and as a result to speed up and improve decision-making processes and hence to support the achievement of organisation goals [3]. As a common side effect, security and privacy have become one of the crucial concerns related to data storage and usage within organisations. this is due to the ethical context, changes in the legal context, the proliferation of cyber criminals, increased malicious insiders, and new attack techniques which have led to the propagation of large scale security breaches in recent years [4, 5]. As reported in [6], about 20.8 billion things will be interconnected around the world in 2020. This in-

creased instrumentation and interconnection will lead to a big rise of cybersecurity issues and safety concerns due to accidental information breaches and organised hacking attempts to various automated systems such as power grid, health, education, banks, government and other private and public systems.

The aforementioned challenges become critical when data governance is not applied in an organisation exploiting big data sets for decision-making. These challenges jointly drive a need to develop a big data governance framework to guide the usage of big data for current decision-making and to ensure the quality and availability of big data for future use. Big data governance involves coordination of people, policies, processes, strategies, standards and technologies to allow organisations to utilise data as one of their critical business assets whilst simultaneously ensuring consistency, usability, integrity, availability, reliability, security and auditability during the entire life cycle of the big data [7].

This paper proposes a framework of big data governance so that big data can be appropriately collected, curated, stored, transmitted, and utilised. Briefly, the proposed framework is guided by governance goals and uses the Evaluate-Direct-Monitor (EDM) cycle model [8] as the governance principle to enhance and support data architecture. Therefore, the framework not only guides the organisations in making better data-driven decisions, but also supports an organisation to efficaciously achieve its organisational outcomes guided by big data. The framework enables organisations to properly manage their data assets and maximise the value of big data, thus to enable and encourage good practice regarding data.

This paper also presents an example implementation of the proposed big data governance framework in the field of cybersecurity. Data analytic tools employing big data and corresponding technologies are increasingly frequently employed to support cybersecurity. Given the wide availability of such data in cyberspace, there are many opportunities to develop and employ such tools. Cybersecurity covers a wide range of aspects of data and network security; and network intrusion detection is one of the common approaches in assuring cybersecurity [9–13]. This paper therefore explores the integration of the proposed big data governance framework in a network intrusion detection system that protects data storage, flow and processing.

The remainder of the paper is organised as follows. Section 2 briefly reviews big data, cybersecurity, and data governance as the underpinnings of this work. Section 3 introduces and discusses the proposed framework of big data governance. Section 4 presents an implementation of the proposed big data governance framework in the field of cybersecurity using a special case of network intrusion detection as example. Section 5 concludes this study and explores possible future directions.

## 2 Background

This section reviews the features of big data, the existing related cybersecurity approaches, and the general data governance frameworks, principles, and lays the foundation for the proposed big data governance framework.

### 2.1 Big Data

Big data is a term associated with massive data sets having larger, more varied and complex structures that are difficult to store, analyse and visualise using traditional data processing technologies [14]. Big data refers not only to the volume of data, but it is also featured by other aspects associated with the collection and utilisation of large volume of data. Big data is commonly described by the 3Vs (i.e., volume, velocity and variety) [15], as demonstrated in Fig. 1.

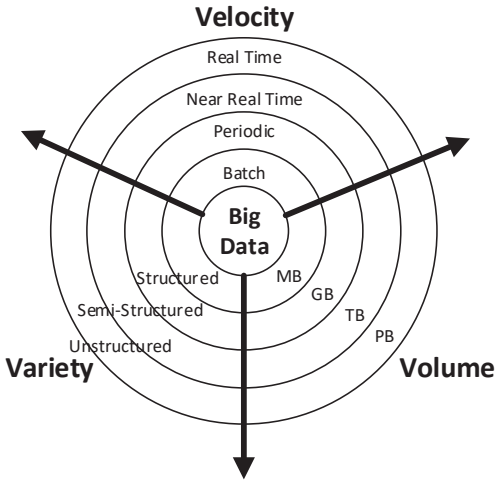


**Fig. 1** The 3Vs of big data - Volume, Variety and Velocity

*Volume:* The most obvious characteristic of big data is the large volume or the huge quantity of data that has been generated and stored, which is usually referred to as the vertical scalability of data. At the moment, it is estimated that 2.5 quintillion bytes of data is created

each day; and there will be an increase of 300 times of data generated every day in 2020 compared to the volume generated in 2005 [16]. This volume usually well exceeds the limitation of traditional column and row rational database, and thus new storage technologies are required to accommodate the big data.

*Variety:* The variety indicates that the data can be sourced from a number of domains with three typical types: structured, semi-structured and unstructured. this variety increases the horizontal scalability of data. Compared with structured data, which usually has already been tagged and can be easily mapped into pre-designed fields, such as tables in a spreadsheet or database, unstructured data appears more random and is more difficult to sort and analyse. Typical examples of unstructured data include emails, videos, photos and audio files. The semi-structured data sets usually do not reside in relational databases or tables, but they do contain tags to separate data elements. The JSON and XML documents commonly belong to this group of data type [16].

*Velocity:* The velocity represents the speed at which the data is generated and needs to be processed in order to meet the demands. Velocity essentially measures how fast the data is created, stored, analysed and visualised. Big data technologies are expected to generate and process data in real-time or near real-time [16], whilst the traditional data handling approaches are only able to deal with data using data snapshots in batches. Therefore, big data requires more powerful data processing mechanism to deal with data streaming in real-time.

The 3Vs have been widely used to describe big data, that is, big data sets are of high-volume, high-variety and high-velocity. In addition, a fourth V of big data, Veracity, has been recently proposed [17]. Veracity refers to the trustworthiness of the data indicating to what extent the data can be confidently used to make crucial decisions by an organisation. Big Data is still a fast evolving area involving very active research and a growing number of applications. As such, unsurprisingly, the definition of big data has also continued to evolve. Nevertheless, most of the definitions are similar to: "Big Data represents the information assets characterised by such a high volume, velocity, variety and veracity to require specific technology and analytical methods for its transformation into value" [17]. Based on this, the generalised definition of Big Data has been extended to include big data processing technologies that realise or extract the value from big data.

### 2.2 Cybersecurity

Following on from the rapid growth of the Internet, more and more devices are being networked to create Internet of Things (IoT). Essentially, a plethora of devices for capturing a wide range of data utilises the ubiquitous connectivity provided by various networks and clouds to share data over the internet. Potentially, such shared data is valuable for organisations if appropriately exploited. In order to protect data sharing in the cyberspace, cybersecurity has become an elevated risk that is amongst the most pressing issues affecting businesses, governments, other organisations, and individuals domestic devices (i.e., televisions, Smart Meters, etc).

Data governance plays an important role in such solutions to not only help organisations understand what data they have to protect but also guide them to achieve their goals, which can be expressed from the following two aspects:

1. Identifying data risk: Personally identifiable information (e.g. contact details) and personal health information (e.g. individual medical records) constitute sensitive data, which could cause reputation and financial risks for the organisations. Data governance tools support the identification of sensitive data [18].
2. Controlling safer access: Data users are not always required to view/access sensitive data for daily usage. It is important to control such sensitive data be only accessed when required and necessary. By correctly applying the data governance tools, privilege data access can be effectively controlled [19].

Traditionally, data-driven network security solutions, such as network intrusion detection system (IDS) and security information and event management (SIEM), are employed to identify anomalies and suspicious activities by analysing transaction logs and network traffic data, and thus provide protection to the organisations from network threats [20]. However, it is becomming more and more difficult for such tools to handle the increasingly larger traffic data sets associated with the use of IoTs for big data collection based on the following two reasons [21, 22]:

1. Traditional techniques were not designed to handle and manage any semi-structured or unstructured data, but this is very common in the big data. It is possible to transform the unstructured data into the structured presentation, to meet the requirement of traditional tools. However, this is an additional and time-consuming process which can be very costly;

2. Traditional techniques are relatively inefficient in storing, retaining, retrieving, accessing and processing a large volume of information implied by big data. Those tools were not integrated with big data technologies.

The issues for traditional tools can be readily solved by applying big data technologies. For instance, the big data tools such as Piglatin scripts and regular expressions, can query data in flexible formats, for both structure and unstructured data. In addition, big data and its distributed systems provide high-performance computing models, which enable the storage and analysis of large heterogeneous data sets at an unprecedented scale and speed. Therefore, the network security issues in association with big data can be targeted by: 1) collecting a massive scale of traffic data; 2) performing deeper analysis on the data; 3) generating and providing network security related models; 4) achieving real-time data analysis of a massive scale of streaming traffic data based on the models.

### 2.3 Data Governance

The digital era provides unprecedented opportunities for the public and private sectors and organisations to collect, store, process and exchange large volumes of data; therefore, they face increasing challenges in data security, data structure management, data privacy preserving and data quality curation. Data governance are the concepts or frameworks that can be used by organisations to address such challenges in managing the processing of digital assets.

Data governance encompasses the people, processes, procedures and technologies to enable an organisation to exploit data as an digital asset [23]. It provides the general framework for the administration and maintenance of the data quality, security, availability, usability, relevancy and integrity. It also ensures that the authentic data is appropriately utilised for setting business goals, maintaining business processes and making critical decisions. Data governance often requires a continuous process to force the cleaning and storage of a large volume of data being generated by an organisation or sourced from third parties. The motivation of applying data governance is to make sure that there is a sustainable means of utilising data to achieve the organisation's business goals and purposes. The information technology (IT) and other business departments must come together to define the rules and strategies that govern the data and define the elements of the data from acquisition, through management and storage, to utilisation and visualisation [24].

Data governance policy establishes the roles and responsibilities of data usage, sets up best practices of data protection plan and ensures that data is properly documented within an organisation [25]. Like any other asset of an organisation, data needs a proper governing policy. Data governance defines the access of the data, the security level of the data, the quality of the data, and the organisation goals on data usage. The data governance policy can be authored by the internal team within an organisation or experts from outside the organisation [25].

Data governance strategy is another key ingredient which defines how information extracted from the data is shared; enforces the culture of using data; reveals the drawbacks that data governance might face and the required budget [26]. More specifically, it articulates who is responsible, accountable and informed regarding the data and how decision will be made from the data. It provides the basis for data management processes to be followed by the entire organisation. It is also an integral part in overcoming data governance limitations and helps to deliver expected business goals and values.

Data governance practices are usually guided by a framework for data collection, management, storage and utilisation. The framework is designed to ensure the confidentiality, usability, accessibility, quality, and integrity of data [26]. It must support data flow and business process within an organisation as well as organisation's culture and structure. It helps to guide staff to perform their roles in data management. A well-established data governance framework usually encompasses data management strategies, corporate drivers, data management structures, technologies and methods [26].

## 3 Big Data Governance

This section discusses data governance challenges and the proposed big data governance framework for addressing the challenges.

### 3.1 The Challenges

Traditional data governance frameworks as reviewed in Section 2.3 usually only consider data principles, data quality, and metadata management for traditional structured and snapshotted reasonably sized data sets, rather than high volume, high variety, and high velocity live data. The large volume of data processed within or outside organisations in a big data environment requires an extra level of management for data quality, security, and the ethical processing of data. In addition, the

combination of big data and business tasks may lead to more frequent and higher level of risks for data breach. The major challenges that the existing data governance frameworks face for big data are summarised in Table 1 [27].

**Table 1** The challenges of existing data governance [27]

| No. | Challenge |
|---|---|
| 1 | Lack of big data governance frameworks. |
| 2 | Shortage of the required skilled people on big data. |
| 3 | Big data security and privacy. |
| 4 | Lack of required tools to generate insight from the data in timely manner. |
| 5 | Organisations resistance to use data in goal setting and decision making. |
| 6 | Insufficient knowledge on big data by the business managers. |
| 7 | Digitisation of the businesses remain a challenge to most of the organisations. |
| 8 | Complexity of the data collected and stored (e.g. unstructured data). |

Data governance in a big data environment mainly focuses on three areas, as listed below:

**1) Data architecture:** Big data is captured from a great variety of sources, which can be structured, semi-structured, or unstructured. Unstructured and semi-structured data are usually random and difficult to be processed. Such structure variety of data increases the difficulty of data management. In addition, different organisations maintain their own standardisation of data structures, which hinders information sharing across systems between different organisations, thus significantly reduces the efficiency the information exploitation and utilisation.

**2) Data quality:** Data quality can be an issue for big data management, as most of the exiting measures cannot be directly applied to big data. For instance, the real-time feature of big data requires organisations to improve data access efficiency, thus to reduce the delay for data transmission. In addition, it may sometimes require the organisations to store the same or conflicting data in different systems, which often leads to synchronisation or conflict resolution issues, and thus reduces data quality accordingly.

**3) Data security, privacy and ethical considerations:** Data sharing and analysis processes can efficiently increase the value of data for an organisation with better data transparency, but it in the same time potentially inappropriately exposes sensitive data or leaves doors to hackers for unauthorised data manipulation. Actually, big data security and privacy are amongst the biggest challenges of data governance in the era of big data. In addition to this, the ethical aspects of data processing has become an increasingly important issue along with the growing concerns of data providers and wider participation of data donations.

### 3.2 Proposed Framework

The existing issues regarding the standardisation of data structures, data quality management and data security and privacy are the main challenges for designing and developing effective big data governance frameworks. This paper presents a big data governance framework from the perspective of big data application and innovation, by considering the goals, data protection enablers, the governance domain, and the principles, as shown in Fig. 2. Briefly, this framework is guided by governance goals in accordance with the organisation's strategic data-related goals, using the EDM cycle model as the governance principle, to effectively enhance data architecture, data quality, and data security and privacy.

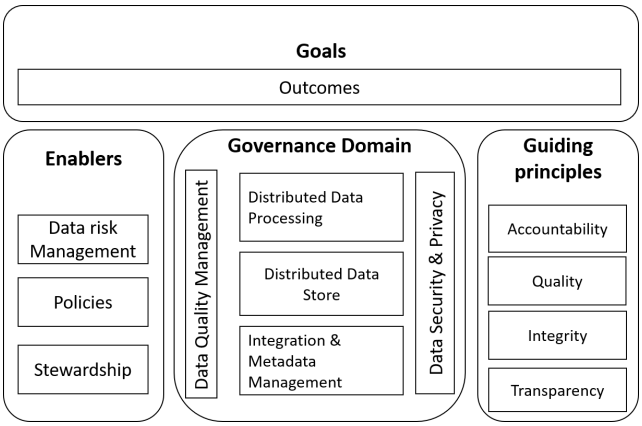


**Fig. 2** Big data governance framework

#### *3.2.1 Goals*

The proposed big data governance framework allows organisations to protect big data whilst making maximum value from the big data, thus to enable and encourage good practices regarding the utilisation of data. In other words, the goals of the framework not only guide the organisations to make better decisions around data usage based on their key performance indicators, but also help organisations efficaciously achieve the outcomes in innovative utilisation of data, which often leads to a culture change of organisations for deep data-driven processes. The goals are usually at the core of those organisation's strategic aims that can be realised by

exploitation of big data using modern information and communication technologies.

### *3.2.2 Enablers*

The enablers define the procedures, rules, and the related enforcement of such procedures and rules through stewardship, which jointly establish the foundational practice standards to limit the risks and maximise the values of the big data and its utilisation. The procedures and rules cover all the key aspects and domains of data processing from data collection, through to storage and transmission, utilisation and finally archiving. For instance, data can leak or be unofficially accessed and manipulated through a variety of ways, from an accidentally lost device or released password to a dedicated organised network attack. The greater diversity of data sources and the storage locations of data in a big data environment make the situation more challenging to handle. In this example, risk prevention and mitigation procedures and rules need to be developed based on such potential risks in an effort to minimise the risks.

In order to ensure the application of the defined rules and procedures in correspondence to the organisation's data handling practices, stewardship takes place in the framework for enforcement, usually by data stewards. The data stewards are the organisation representatives of others who have concerns regarding the processing of data. Data stewards can be in the form of data stewardship council or independent individuals depending on the organisations. The stewardship council also works closely with the procedure and rule makers in providing feedback and insights in the effectiveness and efficiency of the application of procedures and rules; such inputs, in addition to the inputs from data providers, keepers and manipulators, are all taken into consideration during periodical reviews and revisions for data processing procedures and rules.

### *3.2.3 Guiding Principles*

The principles for big data governance support organisations to manage and process big data in providing supplementary guidance for any uncovered aspects specified in data processing procedures and rules as discussed in the last subsection, whilst such procedures and rules define the practical standards of the key aspects of data processing. The procedures and rules are compulsory, which are monitored by the stewardship workgroup. However, the principles discussed in this subsection only suggest good practices, which may not be taken in practice depending on the situations. The guiding principles in this work focus on data accountability, integrity, auditability, and transparency as depicted in Fig. 2. To enable their effectiveness, the guiding principles should be kept simple and understandable.

The organisations should continuously evaluate any change that might happen to the data over a period of time to ensure data integrity which is essential for effective data usage. Data integrity will be achieved by making sure that, data are clearly defined, properly controlled and appropriately accessed within an organisation. As a result of following the principles, data can be better aligned with the organisation strategies as well as cross-business requirements. Data handling procedures need to be transparent to protect an organisation from potential data breach incidents whilst allowing data to be used strategically. Transparency helps to reveal how sensitive data was handled during evaluation so that an internal or third party auditor, or any other data stakeholders, can understand data-related procedures.

In addition, data-related decisions, processes and actions should be auditable, which are supported by appropriate documentation in compliance-based and operational auditing requirements. Accountability generally defines the accessibility of data and the credibility of data operators. In order to facilitate access control to the data, all departments of an organisation need to cooperate to enhance data ownership and accountability. If all departments are accountable and responsible on the data, data breach will be of less concern within an organisation.

### *3.2.4 Governance Domain*

The governance domain describes the data governance objectives that the organisation should focus on whilst conducting data governance activities, which mainly consists of five components, as shown in Fig. 2. Amongst the five components, data quality management and data security and privacy run across all the governance domains, which guarantee the usefulness and effectiveness of big data and the appropriate protection and privacy-preservation of big data during utilisation.

*Data Quality Management:* The use of big data generates data quality issues that are associated with data-in-motion and data-at-rest, as data of poor quality is often inevitably generated and collected, which may increase the negative impacts on organisations operations and decision makings. Data quality management aims to measure, improve and certify the quality and integrity of production, testing, and archival data [28]. Various approaches can be used for big data quality management, in an effort to resolving conflicting data

instances, handling imbalance big data sets, and removing noise amongst others [29, 30].

*Data Security, Privacy and Ethics:* Data security, privacy and other ethics implications form a prime concern when collecting, transmitting, storing, and utilising big data. Big data is often gathered from a great variety of sources, and usually include sensitive information. For instance, the inferred behaviour data, such as work location, buddy list and hangouts, might be classed as private; and in some more serious cases of demographic data, user names, phone numbers and credit card information are very typically used, during data analysis processes. The recently launched EU general data protection regulation (GDPR) is the most important change in data privacy regulation in the past 20 years, which provides the detailed guidelines for organisation in data procession [31]. In this important domain, all the data processing mechanisms are designed based on the GDPR and other policies, the procedures, rules and principles in order to mitigate risks and protect data assets.

*Integration and Metadata Management:* This domain is in the bottom of the governance domain, and can be considered as a data landing zone, which links the data connector and the governance domain. In this domain, multiple methods and tools are often integrated to help understand the data context as well as the contents. Once the contexts and the contents of the data are identified, they are passed to the upper level for the storage. Therefore, big data governance uses integration and metadata management to impose management discipline on the collection and control of data.

*Distributed Data Storage Management:* In the traditional approaches, high-performance computing components, such as the dedicated servers, are utilised for data storage and data replication. Since a huge amount of data are usually generated and collected in a big data environment, those high-performance dedicated servers often fail to meet the performance requirement led by big data. Therefore, this domain aims to provide the methods to allow a large amount of data to be stored and transmitted via a usually distributed architecture, such as Dynamic and Scalable Storage Management (DSSM) [32].

*Distributed Data Processing Management:* With the rapid growth of emerging applications, such as social media and sensor-based networks applications, a variety of data needs to be continuously processed. Obviously, the traditional stand alone solution is no longer suitable for this time limited processes in a live manner. The domain of distributed data processing management then provides a highly efficient framework for big data processing, which allows the analysis of a large amount of data with a reasonable and often acceptable timeline. Typical platforms for such tasks include Spark, MapReduce and Hadoop, amongst other.

### *3.2.5 Confronting New Challenges*

The definition of big data continues to evolve due to the fast changes in the landscape of various supporting digital technologies. Increasingly more-Vs models of the Big Data have been introduced, such as Volume, Variety, Velocity, Veracity, Validity, Value, Variability, Venue, Vocabulary, and Vagueness [33]. These new challenges of big data often require the proposed framework to be scalable to confront such challenges. Generally speaking, the new challenges can usually be grouped into three areas, data challenges, process challenges, and management challenges. In particular, data challenges relate to the characteristics of the data itself; process challenges are usually associated with a set of big data processing techniques; and management challenges cover all the privacy and security issues. Those three components are interlinked to form a data life cycle, as outlined in Fig. 3. Therefore, the proposed Big Data governance framework is extendable and adaptable to handle new challenges of big data, as each individual components in the governance principle is extendable, as also shown in Fig. 3.

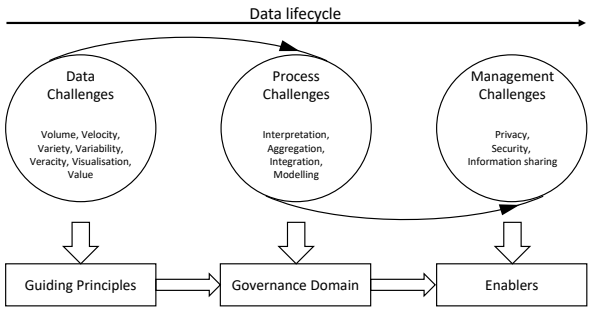


**Fig. 3** Confronting new challenges

## 4 Big Data Governance in Cybersecurity

Big data is often collected from multiple sources with different data capturing equipment, such as IoTs and other specialist devices. The consequence of this is the crucial issues around data security and privacy as such devices have commonly been designed without sufficient consideration on data security. Therefore, cybersecurity has become an increasingly more important and unneglectable research filed to fill such security gap along with the increasingly wider utilisation of big data. This

section discusses the implementation of the big data governance framework introduced in the last section to support the secure and ethical use of big data in this field.

Cybersecurity is the practice of protecting computer and network infrastructures, the operating systems, software programmes run on the infrastructures, and all the data stored or transmitted through the infrastructures from digital attacks and any other misuse, as introduced in Section 2.2. Cybersecurity therefore covers a very wide range of spectrum regarding the hardware and software systems for digital information processing, with the network security as the most common aspect. Furthermore, network intrusion detection is the most common measure to implement network security. Therefore, without losing generality, this paper takes network security as the case in discussing the implementation of the proposed big data governance framework, as illustrated in Fig. 4.

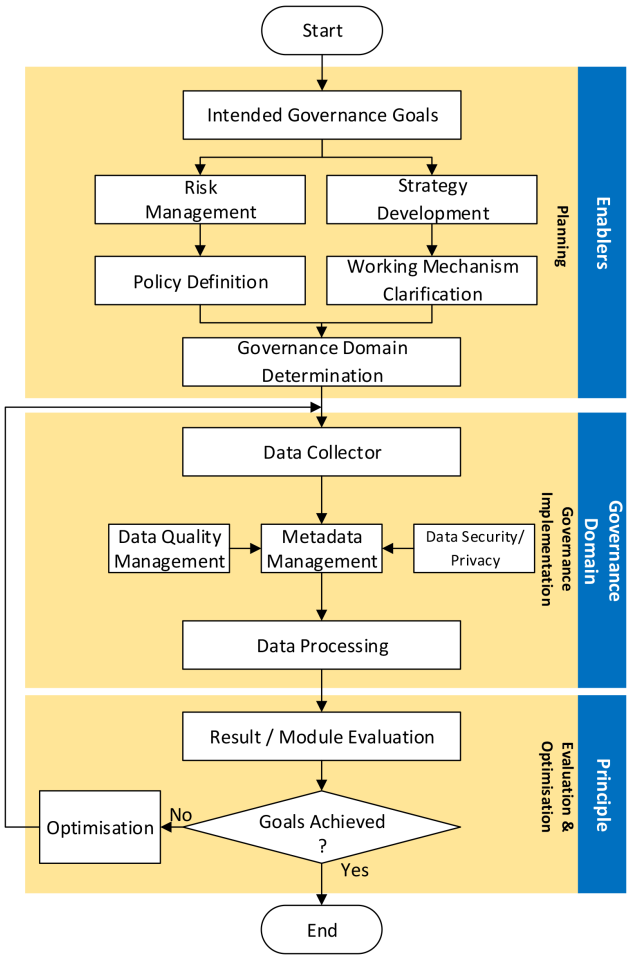


**Fig. 4** Big data governance framework for network security

## 4.1 Planning

The implementation of the proposed data governance framework for cybersecurity needs to be carefully planned to meet the strategic data-related goals in an organisation. In this case, the system needs to be realised by the implementation of the cybersecurity project using big data, to achieve the project and governance goals, which is defined by assessing the current environment of the system and the organisation's key performance indicators. As discussed earlier, this paper uses the intrusion detection system as the running example for the discussion of the proposed approach. Based on this identified goal, possible data risks during the data governance process can be identified; and accordingly, corresponding procedures, rules and principles, such as data storage procedures, data usage principles, health and safety rules, ethical procures can be developed, to address the identified risks. At the same time, the strategies, tools, and mechanisms in implementing the cybersecurity project are also determined and developed.

All the procedures, rules and principles should be implemented in this stage. For the proposed framework, representatives from all departments will establish a data governance council for procedures, rules and principles making as well as monitoring. Take the data access rules as an example in this subsection. Data access rules define the standards and mechanisms for access granting to internal and external users. It is a responsibility of the data governance council to create and grant various access levels of the data in accordance to the needs of various users. The data council team must also work with business partners and data providers to ensure that relevant data are manipulated in align with the predefined rules and regulations by partners. Once the rules and principles are set up, every staff members within an organisation is required to understand the value of data and abide to the regulations that govern the appropriate use of data.

## 4.2 Governance Implementation

The implementation of the data governance framework and the implementation of the network intrusion detection itself were carried out simultaneously. They are discussed jointly in the following subsections.

*Data Collection:* Data collection is the first step in implementing the goal of detecting network intrusions. The quality of the collected data directly affects the performance of the entire system. There are mainly three types of data that can be collected for cyber intrusion

detection: 1) network data packets, which can be collected as full packet capture (FPC), in packet capture (PCAP) format, by applying network packet capture tools, such as Wireshark, TCPdump and TShark, 2) logs of network devices, such as firewall logs, VPN logs and server logs, and 3) event alert information, which are the data generated by firewalls and anti-virus system to alert the network administrators when potential threats are detected.

*Metadata Management:* Once raw data has been collected, either off-line or in-time, it will be passed to the metadata management block for pre-processing, interpreting and labelling. This usually requires a huge storage space. For instance, a 10Tb storage will be required for capturing 1 Gb data stream using the PCAP format for 24 hours [34]. That is equivalent to 900TB storage space for a period of 90 days data collection. In fact, the most interested and useful information for network security analysis is allocated in the packet protocol header, which only occupies 4% of the total size of the PCAP packets. Logs often require less storage space compared to PCAP, but it needs to be structured for data analysis. The metadata management therefore needs to integrate the existing techniques, such as data cleansing and feature extraction tools, to extract the context and content meaning of the captured data for further data analysis. In order to process large big data, distributed data storage and processing are hence required as discussed below.

*Distributed Data Storage and Processing:* Hadoop is a software framework, which is designed to minimise the big data processing time by distributing data storage and processing. In particular, two main components have been provided by Hadoop, Hadoop Distributed File System (HDFS) and MapReduce (a parallel programming model) [35, 36]; the working mechanics of these components are illustrated in Fig. 5. In particular, Hadoop splits the data and distributes them to all the nodes, and runs the applications using the MapReduce algorithm, where the data is processed in parallel, thus to enable the processing of large amount of data which is traditionally impossible [37]. By applying the Hadoop, the collected raw data can be efficiently extracted [37, 38]. The reassembled data set is saved in the Hadoop HDFS again for distributed data utilisation.

*Data Quality Management:* Data quality management is applied to make sure the big data are of appropriate value. For instance, data from different resources may be conflicting to each other, and thus it is important to resolving the conflict before the data used for decision-making. Various approaches are available in the literature for data curation; and in this

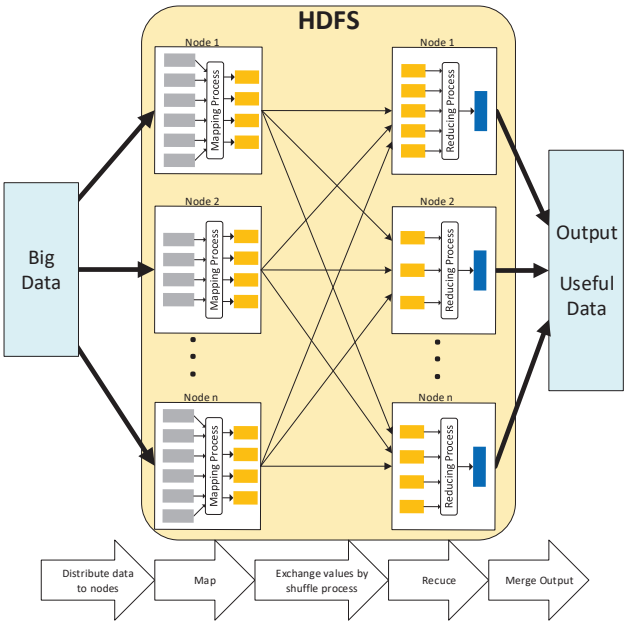


**Fig. 5** The Hadoop Framework

paper, the applied collection framework is employed, which curates the quality and value of big data in four steps, including indicating threads, determining weight for threads, allocating source of data and filtering unnecessary data [34].

*Data Security and Privacy:* All the data in this case, including the raw data and the structured data, needs to be properly secured to prevent any data breaches. As the main equipment for data capturing in this study is the network itself, the captured data therefore can be saved in a separate secured intranet with the support of access level control. The network data may also implies privacy concerns, due to the presence of personal and organisational sensitive data. For instance, the IP addresses could be used to identify individual users or organisations that may collectively provide insights for sensitive user consuming habits and organisational commercial information. Therefore, privacy-preserving mechanism must also be applied, in addition to any other aspects of data protection based on the GDPR.

### 4.3 Evaluation and Optimisation

The collected data, after pre-processing, can then be fed into artificial intelligence or machine learning approaches for intrusion detection. In order to reduce noise and improve accuracy, feature selection approaches, such as the work reported in [39], may be applied first, depending on the nature of the data set. A large number of machine learning approaches have been proposed for network intrusion detection, such as [10, 11, 40],

and one of such approaches can be applied, for a given problem, subject to its performance. The velocity and volume of big data often lead to imbalance, sparsity and evolving nature of the dataset, various adaptive approaches, such as adaptive fuzzy interpolation [41, 42], can therefore be applied to handle such situations. From this, the results can be evaluated, and the approach itself can then be optimised if required, in order to optimise the approach and thus maximise the achievement towards the goals.

## 5 Conclusion

This paper presents a big data governance framework to support organisations to appropriately manipulate both structured and unstructured big data, make maximum value from the big data, and enable and encourage good practice regarding big data. The framework is proposed to support organisations to make better business decisions, whilst helping organisations to efficaciously achieve data security, usability and availability. An implementation of the framework based on a case of network security is also presented in this paper. This case study illustrates how to safeguard data when implementing network security.

Although promising, the work can be improved in multiple directions. Firstly, it is worthwhile to effectively validate and evaluate the proposed framework by implementing the framework in a real-world network environment. Also, the proposed work is only presented based on a particular case, and thus it would be very appealing to extend the work to other big-data-based cybersecurity cases. In addition, it is interesting to systematically compare the proposed framework with the existing data governance frameworks for traditional data sets. Finally, it is worthwhile to consider how traditional model governance approaches, such as the one reported in [43], could be extended to support models based on big data.

## References


1. Yang L, Li J, Chao F, Hackney P, Flanagan M (2018) Job shop planning and scheduling for manufacturers with manual operations. Expert Systems DOI 10.1111/exsy.12315
2. Tsai C-W, Lai C-F, Chao H-C, Vasilakos A V (2015) Big data analytics: a survey. Journal of Big data 2(1):21
3. Chen J, Chen Y, Du X, Li C, Lu J, Zhao S, Zhou X (2013) Big data challenge: a data management perspective. Frontiers of Computer Science 7(2):157–164
4. Terzi D S, Terzi R, Sagiroglu S (2015) A survey on security and privacy issues in big data. In: Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference for, IEEE, pp. 202–207
5. Singh M, Halgamuge M N, Ekici G, Jayasekara C S (2018) A review on security and privacy challenges of big data. In: Cognitive Computing for Big Data Systems Over IoT, Springer, pp. 175–200
6. Gartner (2015) 20.8 Billion will be Connected by 2020. "`https://www.gartner.com/newsroom/id/3165317/`, accessed: 2018-11-14
7. Morabito V (2015) Big data governance. In: Big data and analytics, Springer, pp. 83–104
8. Calder A (2008) ISO/IEC 38500: the IT governance standard. IT Governance Ltd
9. Li J, Qu Y, Chao F, Shum H P H, Ho E S L, Yang L (2019) Machine Learning Algorithms for Network Intrusion Detection, Springer International Publishing, Cham, pp. 151–179
10. Bostani H, Sheikhan M (2017) Modification of supervised opf-based intrusion detection systems using unsupervised learning and social network concept. Pattern Recognition 62:56–72
11. Yang L, Li J, Fehringer G, Barraclough P, Sexton G, Cao Y (2017) Intrusion detection system by fuzzy interpolation. In: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6
12. Naik N, Jenkins P, Kerby B, Sloane J, Yang L (2018) Fuzzy logic aided intelligent threat detection in cisco adaptive security appliance 5500 series firewalls. In: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8
13. Li J, Yang L, Qu Y, Sexton G (2018) An extended takagi–sugeno–kang inference system (TSK+) with fuzzy interpolation and its rule base generation. Soft Computing 22(10):3155–3170
14. Sagiroglu S, Sinanc D (2013) Big data: A review. In: 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47
15. Mark B (2011) Gartner says solving 'big data' challenge involves more than just managing volumes of data
16. Madden S (2012) From databases to big data. IEEE Internet Computing 16(3):4–6, DOI 10.1109/MIC.2012.50
17. Mauro A D, Greco M, Grimaldi M (2016) A formal definition of big data based on its essential features. Library Review 65(3):122–135

20. Miller D, et al. (2011) Security information and event management (SIEM) implementation. McGraw-Hill,
21. Cárdenas A A, Manadhata P K, Rajan S (2013) Big data analytics for security intelligence. University of Texas at Dallas@ Cloud Security Alliance pp. 1–22
22. Zikopoulos P, Eaton C, et al. (2011) Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media
23. Khatri V, Brown C V (2010) Designing data governance. Communications of the ACM 53(1):148–152
24. Tallon P P (2013) Corporate governance of big data: Perspectives on value, risk, and cost. Computer 46(6):32–38
25. Rosenbaum S (2010) Data governance and stewardship: designing data stewardship entities and advancing data access. Health services research 45(5p2):1442–1455
26. Berson A, Dubov L, Plagman B K, Raskas P (2011) Master data management and data governance. McGraw-Hill
27. Katal A, Wazid M, Goudar R (2013) Big data: issues, challenges, tools and good practices. In: Contemporary Computing (IC3), 2013 Sixth International Conference on, IEEE, pp. 404–409
28. Soares S (2012) Big data governance: An emerging imperative. Mc Press
29. Yang L, Neagu D, Cronin M T D, Hewitt M, Enoch S J, Madden J C, Przybylak K (2013) Towards a fuzzy expert system on toxicological data quality assessment. Molecular Informatics 32(1):65–78
30. Pipino L L, Lee Y W, Wang R Y (2002) Data quality assessment. Commun ACM 45(4):211–218
31. (2018) General data protection regulation. `https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en`, accessed: 2018-11-28
32. Kumar A, Bawa S (2012) Distributed and big data storage management in grid computing. arXiv preprint arXiv:12072867
33. Moorthy J, Lahiri R, Biswas N, Sanyal D, Ranjan J, Nanath K, Ghosh P (2015) Big data: Prospects and challenges. Vikalpa 40(1):74–96
34. Sanders C, Smith J (2013) Applied network security monitoring: collection, detection, and analysis. Elsevier
35. Ramesh S, Rauf H A, Victor S (2017) Development of hybrid intrusion detection system on big data for detecting unknown attacks by using ahsvn. International Journal of Technology in Computer Science & Engineering 4(2)
36. Shvachko K, Kuang H, Radia S, Chansler R (2010) The hadoop distributed file system. In: Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on, Ieee, pp. 1–10
37. Liu X, Song B (2016) Hadoop-based mass data tcp packet reassembly technology. Computer Engineering 42(10):113
38. Mavridis I, Karatza H (2017) Performance evaluation of cloud-based log file analysis with apache hadoop and apache spark. Journal of Systems and Software 125(Supplement C):133 – 151
39. Sun Q, Qu Y, Deng A, Yang L (2017) Fuzzy-rough feature selection based on $\lambda$-partition differentiation entropy. In: 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 1222–1227
40. Wang G, Hao J, Ma J, Huang L (2010) A new approach to intrusion detection using artificial neural networks and fuzzy clustering. Expert Systems with Applications 37(9):6225–6232
41. Yang L, Chao F, Shen Q (2017) Generalized adaptive fuzzy rule interpolation. IEEE Transactions on Fuzzy Systems 25(4):839–853
42. Yang L, Shen Q (2011) Adaptive fuzzy interpolation. IEEE Transactions on Fuzzy Systems 19(6):1107–1126
43. Palczewska A, Fu X, Trundle P, Yang L, Neagu D, Ridley M, Travis K (2013) Towards model governance in predictive toxicology. International Journal of Information Management 33(3):567 – 582