



Review

Security and Privacy in Big Data Life Cycle: A Survey and Open Challenges

Jahoon Koo , Giluk Kang and Young-Gab Kim * 

Department of Computer and Information Security and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Korea; sigmao91@sju.ac.kr (J.K.); giluk1027@sju.ac.kr (G.K.)

* Correspondence: alwaysgabi@sejong.ac.kr

Received: 26 November 2020; Accepted: 16 December 2020; Published: 17 December 2020



Abstract: The use of big data in various fields has led to a rapid increase in a wide variety of data resources, and various data analysis technologies such as standardized data mining and statistical analysis techniques are accelerating the continuous expansion of the big data market. An important characteristic of big data is that data from various sources have life cycles from collection to destruction, and new information can be derived through analysis, combination, and utilization. However, each phase of the life cycle presents data security and reliability issues, making the protection of personally identifiable information a critical objective. In particular, user tendencies can be analyzed using various big data analytics, and this information leads to the invasion of personal privacy. Therefore, this paper identifies threats and security issues that occur in the life cycle of big data by confirming the current standards developed by international standardization organizations and analyzing related studies. In addition, we divide a big data life cycle into five phases (i.e., collection, storage, analytics, utilization, and destruction), and define the security taxonomy of the big data life cycle based on the identified threats and security issues.

Keywords: big data; life cycle; big data security; privacy

1. Introduction

Recently, big data has garnered considerable attention from the industry, scientific and technology communities, media, and several government departments. Many countries are also using big data to provide services in various fields such as healthcare, medicine, public sector undertakings, distribution, marketing, and manufacturing. Big data is essentially an information-based technology that analyzes large amounts of data to extract valuable information and predicts changes based on the extracted knowledge. It is considered a new source of energy that drives business and technological innovations as well as economic growth. Many economic and political interests drive big data, especially the processes of data integration, analysis, and data mining. In particular, organized big data collected from various sources, such as social media platforms, websites, and global positioning systems will help to identify various socio-economic problems and also help in providing effective solutions and measures. The use of big data in various fields has led to a rapid increase in a wide variety of data resources, and various data analysis technologies, such as standardized data mining and statistical analysis techniques, are accelerating the continuous expansion of the big data market. An important characteristic of big data is that data from various sources have life cycles from collection to destruction, and new information can be derived through analysis, combination, and utilization.

As mentioned previously, big data offers many advantages and potentials for innovation in various fields but also presents many issues and challenges. First, data security, privacy-preserving, and ethical issues are major open challenges in the big data innovation ecosystem and include information management methods, protection of personal or fatal information, and misuse of data

analyses. In particular, a large amount of shared information, including privacy, can be exploited in an interconnected open environment. Hence, various standardization organizations have published related standards for security and privacy-preserving of big data, and privacy protection laws such as the general data protection regulation (GDPR) in Europe and the California consumer privacy act (CCPA) in the United States have been enacted. However, the standards related to big data security only explain the security requirements and lack any description related to security techniques. Furthermore, since the GDPR and CCPA are targeting specific regions, they are not generalized to various organizations and researchers that utilize big data. Second, each phase of the life cycle has data security and reliability issues, and the protection of personally identifiable information is crucial. In particular, user tendencies can be analyzed using various big data analytics, leading to the invasion of personal privacy. Various technologies for preserving security and privacy in a big data environment have been proposed and have been under development until recently. These can be divided and grouped according to the phases of the big data life cycle.

Therefore, this paper identifies threats and security issues that occur in the big data life cycle by confirming the current standards developed by international standards organizations and analyzing related studies. In addition, we divided a big data life cycle into five phases (i.e., collection, storage, analytics, utilization, and destruction), and defined the security taxonomy of the big data life cycle based on the identified threats and security issues. The contributions of this paper include:

1. Analysis of the development status of standardization organizations and studies related to big data security and privacy-preserving
2. Description of the security techniques for each phase according to the threats in the big data life cycle, and writing the taxonomy of security and privacy issues based on related studies
3. Evaluation comparing our proposal with existing big data security and privacy-preserving survey studies

The remainder of this paper is organized as follows. Section 2 analyzes background knowledge and summarizes standards related to big data security and privacy-preserving. Section 3 proposes security taxonomy and describes security technologies based on a big data life cycle. Section 4 compares and evaluates our proposal with the current survey studies. Section 5 summarizes and concludes this paper and discusses future works. In addition, the list of abbreviations used in this paper is presented in Table 1.

Table 1. List of abbreviations.

Abbreviation	Description
ABE	Attribute-based encryption
APK	Android application package
BSI	British standards institution
CCPA	California consumer privacy act
DBSCAN	Density-based spatial clustering of applications with noise
EEA	European economic area
EU	European union
FPE	Format-preserving encryption
GDPR	General data protection regulation
Hadoop	High-availability distributed object-oriented platform
HMAC	Hash-based message authentication code
IBE	Identity-based encryption
IEEE-SA	Institute of electrical and electronics engineers standards association
IoT	Internet of things
ISO	International organization for standardization
ISO/IEC JTC1	International organization for standardization/international electrotechnical commission joint technical committee 1
ITU-T	International telecommunication union telecommunication standardization sector

Table 1. Cont.

Abbreviation	Description
JAR	Java archive
LLOA	Least lion optimization algorithm
MAC	Message Authentication Code
MCA	Multiple correspondence analysis
NIST	National institute of standards and technology
OAuth 2.0	Open Authorization 2.0
Open API	Open application programming interface
OTP	One-time password
PCA	Principal component analysis
PG	Project group
PII	Personally identifiable information
PPDM	Privacy-preserving data mining
PPDP	Privacy-preserving data publishing
PRE	Proxy re-encryption
SAC	Standardization administration of china
SG	Study group
SHA	Secure hash algorithm
SMC	Secure multiparty computation
STC	Special technical committee
TC	Technical committee
TM Forum	Tele management forum
TTA	Telecommunications technology association
TTP	Trusted third-party
WD	Working draft
WG	Working group
Zip-code	Zone improvement plan-code

2. Background and Standards

In this section, we provide background studies related to big data security and privacy issues. Section 2.1 describes the big data life cycle and analyzes threats. Section 2.2 describes the status of the development of standards by various standardization organizations.

2.1. Big Data Life Cycle

This section describes the big data life cycle by dividing it into five phases (i.e., collection, storage, analytics, utilization, and destruction) as shown in Figure 1. In addition, we identify security and privacy threats that arise in each phase.



Figure 1. Big data life cycle.

2.1.1. Collection

In the data collection phase, data is collected from diverse sources, with different formats, such as structured, semi-structured, and unstructured. Ideally, securing big data technology should apply to the collection phase of the life cycle on a preferential basis. It is important to acquire reliable data to ensure that this phase is appropriately secured and protected. Furthermore, additional security measures are necessary to keep data from being released. Some security measures can be used in this phase, such as limited access control and encryption of some data fields. In addition, data can be collected through software, social media, and the internet, regardless of consent from the data provider.

In other words, the data collector may infringe upon the provider's data sovereignty by inappropriately collecting data without any consent. In particular, many people provide implied consent and data voluntarily in the process of performing daily activities such as social media and shopping, and this lack of awareness of consenting without fully understanding the potential ramifications of providing privacy is a major issue. There is also the possibility of acquiring sensitive data through various attacks (e.g., spoofing, phishing, and spamming) that trick or attack providers and collectors.

2.1.2. Storage

In the data storage phase, the collected data is stored for use in the next phase (i.e., data analytics phase). As the collected data may contain sensitive information, it is important to apply efficient precautions for data storage. The stored data needs to be protected from multiple threats by combining physical security techniques and data protection technologies. In cases where it is not completely reliable, such as in the cloud, data integrity and confidentiality must be maintained through privacy-preserving technologies (e.g., encryption and masking). Because the size of data is enormous, data storage services need to be adhered to a distributed storage, and sensitive data must be provided only to authorized persons through access control. In addition, if sensitive data is unintentionally passed beyond consent during the collection process, it must be immediately destroyed.

2.1.3. Analytics

After data collection and storage, the data is processed and analyzed to generate useful knowledge. Various data mining techniques are used in this step, such as clustering, classification, and link rule mining. In the analytics phase, it is important to provide a secure environment for processing and analysis. Data miners can identify sensitive data through powerful mining algorithms and make their systems vulnerable to the invasion of privacy. Therefore, the data mining process and analysis results should be protected from mining-based attacks, and only authorized persons should be allowed. In addition, in the process of analyzing data, the efficiency of privacy protection is inversely proportional to data processing, i.e., it is difficult to increase processing efficiency while protecting sensitive data. Because this is a critical issue, various mining techniques and de-identification techniques to protect privacy are being developed. However, there are issues such as re-identification. The main attacks in the analysis process are as follows: (1) linking attacks allow de-identified data to be re-identified by associating de-identified data with other data. (2) homogeneity attacks re-identify data using the information of homogeneous aggregated data in k-anonymity. (3) background knowledge attacks re-identify data de-identified as k-anonymity based on background knowledge. (4) skewness attacks re-identify data based on the value of de-identified data in a data set that has been de-identified as l-diversity. In addition, because some mining techniques can identify a specific person or extract sensitive data at any time according to the intention of the data miner, and use it for unauthorized purposes, it is necessary to ensure that only some approved miners perform the work.

2.1.4. Utilization

The analytics phase delivers new information and valuable insights that are used by decision-makers. This knowledge is considered sensitive information, especially in a competitive environment. Organizations, in competing against their business rivals, typically take particular care of such valuably sensitive information. Further, they actively ensure the sensitive client personal data is not publicly released. Essentially, to create new information through a combination of analytics of sensitive information is the main purpose of the utilization phase. Even if there is no sensitive data, linking the data collected from various fields can help in identifying a specific person or inferring sensitive data, and this information can be used for other purposes without consent. Furthermore, there is a possibility that sensitive data may be unintentionally inferred through the result of mining, and advertising information that does not agree can be transmitted by identifying a specific target. In addition, there is a possibility that

decision-makers may share sensitive data with third parties to pursue business interests, and require post-processing techniques and audits.

2.1.5. Destruction

In the data destruction phase, data used for analysis is deleted. Basically, privacy data should be destroyed without delay after exceeding the data retention period, unless otherwise specified in other laws and regulations. In addition, data must be destroyed if it is no longer necessary for the intended purpose, or if the data provider withdraws consent. As described previously, there are solutions for physically destroying a hard disk, or other storage devices, for data destruction, and software such as overwriting is used several times. However, the methods involve the disposal of the entire physical/logical space that stores data, making it difficult to delete only some of the data. It can also be difficult to verify the effectiveness of the disposal. In general, it is necessary to destruct the data according to the purpose of the data and the user's withdrawal of consent. However, some organizations use the data despite achieving their intended purpose and withdrawing their consent. More notably, privacy protection problems also arise due to the act of selling data to third-party firms. In addition, due to the nature of big data architectures, there is a possibility that the data cannot be deleted because it is destroyed in a distributed environment.

2.2. Application Scenario Based on Big Data Life Cycle

Generally, there can be five types of users (i.e., data providers, data collector, storage admin, analyst, and decision-maker) and each big data platform in the application scenario, as shown in Figure 2. We briefly describe the user role in each phase as follows.

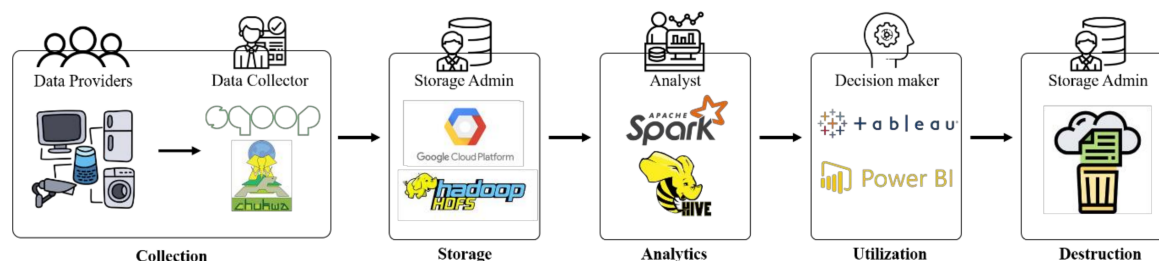


Figure 2. A simple scenario of the application.

A data collector collects data from data providers through various routes such as IoT devices and social network services. As this process may contain sensitive information, appropriate data correction and security measures are essential. A storage administrator stores data from data providers through the cloud system and distributed storage systems. As the storage phase can contain sensitive information from data providers, a storage administrator should use various security techniques to manage them safely. In addition, data erasure should be performed with the statute to ensure the rights of the data provider. An analyst can analyze the data in the repository to obtain appropriate analysis results. However, various privacy issues can arise during mining and analysis, so an analyst should balance data's usefulness and privacy using privacy protection techniques. A decision-maker can visualize and utilize the analyzed results in a variety of ways. Therefore, the utilization phase needs privacy protection techniques, as it can lead to unintentional privacy leakage.

2.3. Standards

This section identifies existing published standards and current developing standards. We describe the de jure and de facto standards that address big data. The organizations of de jure standards include the ISO, ITU, ISO/IEC JTC1, NIST, SAC, and BSI. The organizations of the de facto standard include the TTA, TM Forum, IEEE-SA, and Apache. The standards related to big data security addressed by these

standardization organizations are shown in Figure 3, and an overview of each standard organization is presented in the subsequent sections.

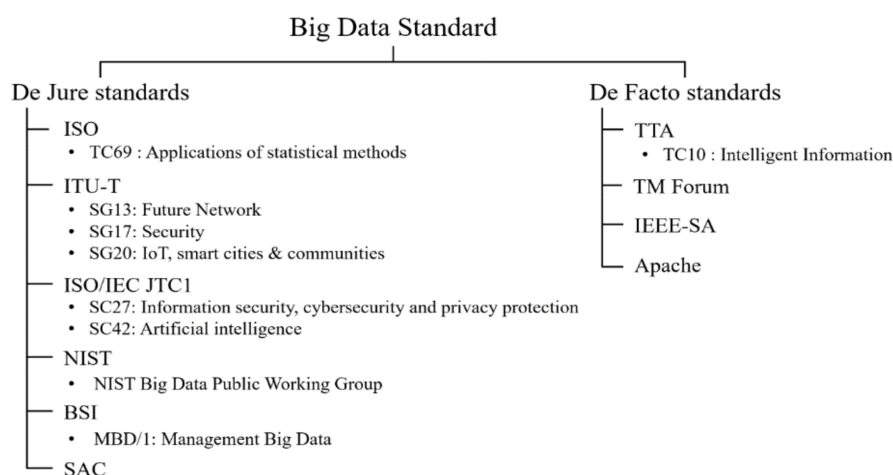


Figure 3. Taxonomy of Standards for Big Data Security and Privacy-Preserving.

The ISO is an international standardization organization comprising representatives of various national standards bodies in 1947 and was established to deal with problems that might arise with different industrial and commercial standards in different countries [1]. The ITU is a unit of the UN and is the organization responsible for all matters related to telecommunications. There are three divisions of which ITU-T is in charge of the standards related to telecommunications [2]. The ISO/IEC JTC1 is the first joint technical committee established in 1987 as a joint venture between the standards body ISO and IEC to designate mutually cooperative standardization of information technology, and several standards are currently being developed by 20 subcommittees [3]. NIST was established in 1901 as the agency responsible for the development of national standardization under the US department of commerce. It was created to improve standards of quality and technology to improve the quality of life and enhance American industrial competitiveness [4]. SAC is a national standardization agency established in 2001 to manage, supervise, and coordinate overall standardization in China, and to promote national interests in the field of international standardization [5]. BSI is the world's first national standards body, established in 1901, responsible for the establishment of various national standards across industry in the UK [6]. TTA is an organization established in 1988 that tests standardization activities and standard products in the field of information and communication in Korea and is the only organization that designates information and communication organization standards in Korea [7]. TM Forum is a non-profit organization established in 1988 by eight companies to solve system and operation management problems through the open systems interconnection protocol. Currently, it provides practical information and technology to help businesses with various service providers in over 180 countries [8]. IEEE-SA is an association under IEEE, an association for electrical/electronic engineering, and related studies established in the US. It is an association that develops global standards for various industries such as electrical, electronic, communication, and medical engineering [9]. The ASF is a non-profit organization that supports open-source Apache software projects. In these projects, the Apache Hadoop project developed open-source software for reliable, scalable, and distributed computing. Hadoop is used by various companies, and in particular, data analytics companies are leveraging it to build platforms [10].

2.3.1. De Jure Standards

De jure is standards established by established standardization organizations (e.g., international standardization organizations, national standardization organizations, and collective standardization organizations) through certain procedures and deliberations. However, it takes 3 to 6 years for

the international standardization organization to establish a standard, and in a field with rapid technological progress, such as information and communication, these standardization activities are often unable to keep up with the market trend. This section describes the de jure standards related to big data, and in particular, Table 2 represents the items related to big data security.

Table 2. De Jure Standards for Big Data Security and Privacy-Preserving.

Affiliation	Number	Title	Limitation	Status
ITU-T/SG 13	X.1147	Security requirements and framework for big data analytics in mobile Internet services	A brief explanation of security requirements	Published
	X.1750	Guidelines on security of big data as a service for Big Data Service Providers	It cannot be viewed	Pre-published
	X.1751	Security guidelines on big data lifecycle management for telecommunication operators	It cannot be viewed	Pre-published
ISO/IEC JTC1 SC 27	20547-4:2020	Information technology—Big data reference architecture—Part 4: Security and privacy	Guideline for functional components not including detailed techniques	Published
	WD 27045.5	Information technology—Big data security and privacy—Processes	It cannot be viewed	Preparatory
	27046.2	Information technology—Big data security and privacy—Implementation guidelines	It cannot be viewed	Preparatory
NIST	SP 1500-4r2	NIST Big data interoperability framework: Volume 4, Big data security and privacy	Architectural security and privacy issues not including the big data life cycle	Published
SAC	GB/T 35274-2017	Information security technology—Security capability requirements for big data services	A rough description of the requirements and	Published
	GB/T 37973-2019	Information security technology—Big data security management guide	insufficient description of the techniques	Published

International Organization for Standardization (ISO)

The organization that conducts big data-related research is referred to as the TC 69—Applications of statistical methods. TC 69 is responsible for standardizing the application of statistical methods, including data generation, collection (i.e., planning and design), analysis, presentation, and interpretation. Although standardization of big data-based statistical analysis is being conducted, standards related to privacy-preserving, which must be dealt with in the analysis process, are not in progress.

International Telecommunication Union Telecommunication Standardization Sector (ITU-T)

Various standards are currently being developed by 11 study groups under ITU-T. Various data standards are being developed in the SG 13 *Future networks*, SG17 *Security*, and SG20 *IoT, smart cities & communities*. SG13 is a study group developing standards related to next-generation networks and cloud computing. Starting with the development of *Requirements and capabilities for cloud computing-based big data* in 2013, they are developing several standards for architecture and requirements related to big data. Recently, various cloud-based big data standardizations, such as the Y.3519 *Cloud computing-functional architecture of big data as a service* and Y.3601 *Big data framework and requirements for data exchange* have been proposed. SG17 is a study group that develops security standards across all fields of telecommunications. Various other big data infrastructure security standards are being developed, such as the X.1147 *Security requirements and framework for big data analytics in mobile internet services*, X.1750 *Guidelines on security of big data as a service for big data service providers*, and X.1751 *Security*

guidelines on big data lifecycle management for telecommunication operators. X.1147 describes the threat in mobile internet big data analysis services and analyzes the security requirements of big data analytics. However, only a brief explanation is presented, and no clear technical proposal is available. X.1750 and X.1751 cannot be viewed, due to pre-publication. SG20 is a study group that develops standards related to the IoT and smart cities. Publishing of Y.4114 *Specific requirements and capabilities of the IoT for big data* have been identified, and research on the use of big data in the IoT environment is in progress.

ISO/IEC JTC1

The WG 9 *Big data* was established to develop standards related to big data in 2015, and standards are currently being developed by SC 27 *Information security, cybersecurity and privacy protection*, and SC42 *Artificial intelligence*. SC27 is a subcommittee established to develop standards related to information and communications technology protection and is developing standards related to big data security such as the 20547-4:2020 *Information technology—Big data reference architecture—Part 4: Security and privacy*, WD 27045.5 *Information technology—Big data security and privacy—Processes*, and 27046.2 *Information technology—Big data security and privacy—Implementation guidelines*. 20547-4:2020 part 4 specifies the security and privacy aspects applicable to the reference architecture, including the big data roles, activities, and functional components. In addition, it also provides guidance on security and privacy operations for big data. This guideline includes functional components for security and privacy preservation. However, it does not describe the required techniques for each functional component. WD 27045.5 and 27046.2 are in the preparatory state and cannot be viewed. SC42 is a subcommittee established in 2017 for the development of AI-related standards, and WG9, which was in charge of big data, was established as the subcommittee, and was incorporated into WG2 *Data*. WG2 is developing reference architectures and frameworks related to big data such as ISO/IEC 20,547 *Information technology—Big data—Reference architecture* and ISO/IEC 24,668 *Information technology—Artificial intelligence—Process management framework for big data analytics*.

National Institute of Standards and Technology (NIST)

Big data-related work is performed by the *NIST Big data public working group*, whereas the *NIST Big data security & privacy subgroup* address big data security. The currently established security standards are the NIST SP 1500-4r2 *NIST Big data interoperability framework Volume 4, security, and privacy*. NIST SP 1500-4r2 presents the limitations of existing security solutions and 15 problems caused by big data. In addition, it identifies privacy and security issues and maps necessary factors and responsibilities according to roles in the *NIST Big data reference architecture* based on it.

Standardization Administration of China (SAC)

Currently, various standards for big data security are established and under development. The standards related to big data security are the GB/T 35274-2017 *Information security technology—Security capability requirements for big data services* and GB/T 37973-2019 *Information security technology—Big data security management guide*. GB/T 35274-2017 specifies that big data service providers should have the organization related to basic security capabilities and data life cycle-related data security capabilities. This standard describes security requirements according to the data life cycle (i.e., acquisition, transmission, storage, processing, exchange, and destruction). GB/T 37973-2019 also describes security requirements and big data security risks, such as the identification of threats and vulnerabilities. However, only a rough description of the requirements exists, and detailed technical statements and necessities are insufficient.

British Standards Institution (BSI)

MBD/1—*Big data—Management big data* is responsible for standardizing big data. BS 10102-1:2020—*Big data. Guidance on data-driven organizations* and BS 10102-2:2020—*Big data. Guidance on data-intensive projects*

is established as a standard. BS 10102-1:2020 addresses the security-by-design and privacy-by-design according to the data life cycle including acquisition, storage, usage, sharing, and destruction.

2.3.2. De Facto Standards

De facto is a standard established by companies and organizations of a specific field. It influences the market economy due to its popularity, and its status is continuously strengthening. This section describes the de facto standards related to big data. However, de facto standards do not address the items related to big data security.

Telecommunications Technology Association (TTA)

Standards related to big data were developed with the establishment of the STC 2 *Cloud/big data special technical committee* in 2014. It has been incorporated into the PG 1004 *Big data* under TC 10: *Intelligent information-based technical committee* and is in progress to standardize activities related to big data such as TTA.KO-10.0900 *Deployment and utilization guidelines for big data according to the data life cycle*. However, standards related to big data security are not in progress.

Tele Management Forum (TM Forum)

Currently, standards related to big data analysis are being developed and established in various ways such as GB979 *Big data analytics guidebook R16.5.1*, GB979D *Big data analytics big data repository R18.5.1*, and TR261 *Data governance functions and implementation R16.0.1*. However, standards related to security have not been addressed.

Institute of Electrical and Electronics Engineers Standards Association (IEEE-SA)

IEEE-SA is developing big data standards for various industries, such as three-dimensional and medical data. Standards related to big data privacy preservation are also being developed through projects such as the IEEE P7002 *Data privacy process* and IEEE P7006 *Personal data AI agent*. IEEE P7002 defines requirements for a systems/software engineering process for privacy-oriented considerations regarding products, services, and systems utilizing employees, customers, or the personal data of other external users. IEEE P7006 describes the technical elements required to create and grant access to a personalized AI that will comprise inputs, learning, ethics, rules, and values controlled by individuals.

Apache

Apache proposed an ecosystem through interoperability with several open-source projects based on the Hadoop, a distributed processing platform for big data analysis. Among them, the yet another resource negotiator and Hadoop distributed file system projects provide security such as authentication and authorization.

2.3.3. Outlook and Drawback of Current Standards

Various standardization organizations are developing standards related to big data. However, it takes a lot of time for a standard to be published, and the gaps during that period create several issues. The important point is despite security and privacy are critical issues in big data, many standard organizations are still being developed or not. In addition, the published standards only present out-of-date technologies as requirements, and there are no detailed descriptions of the technologies. Therefore, de jure and de facto standardization organizations need to publish together to quickly standards for big data security and privacy suitable for the market.

3. Security and Privacy in Big Data Life Cycle

This section defines the security taxonomy for the big data life cycle, as shown in Figure 4. We define the security taxonomy based on the identified threats, security, and privacy issues in the big

data life cycle. Security taxonomy includes technologies required in each phase of the big data life cycle. We describe each security technique and summarize related studies.

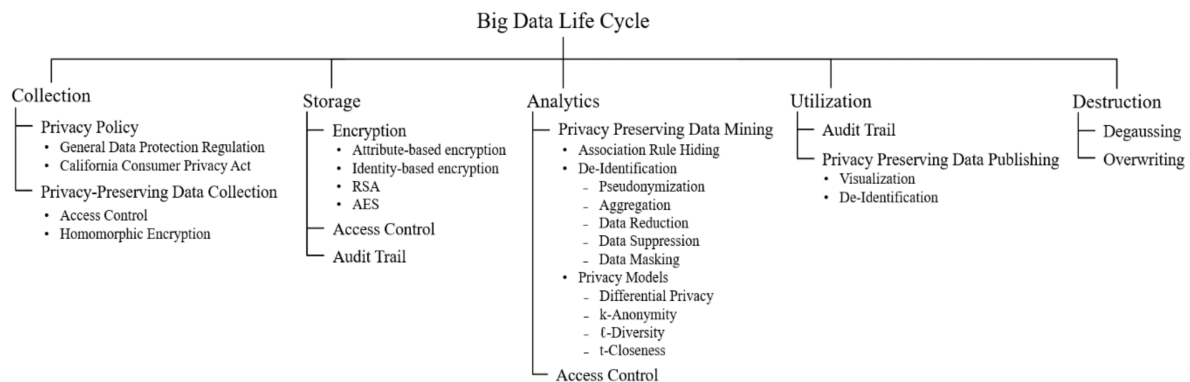


Figure 4. Taxonomy of security and privacy-preserving in the big data life cycle.

3.1. Collection

The collection phase includes privacy policy and privacy-preserving data collection, and Table 3 summarizes related studies of data collection such as the approach and description.

Table 3. Description of conceptual approaches for big data collection.

Type	Papers	Approaches	Descriptions
3.1.1 Privacy Policy	Greene et al. [11]	GDPR	Identify the GDPR concepts and principles and how they can impact the work of data scientists and researchers.
	Stallings et al. [12]	CCPA	Describes how it deals with obfuscation algorithms that can protect privacy.
	Kanika et al. [13]	GDPR and CCPA	Describes the laws dealing with privacy protection when the information provider withdrawal of consent.
3.1.2 Privacy-Preserving Data Collection	Hornyack et al. [14]	Access control	Block unnecessary access and privacy-preserving using shadow data.
	Zhao et al. [15]	Authentication/Authorization	Propose a personal data cloud to store collected personal data and control access.
	Gao et al. [16]	Homomorphic encryption	A PPAS that enables data providers to sell data securely through one-time pad and homomorphic encryption.
	Mittal et al. [17]	Homomorphic encryption	Maintaining accuracy and preserving-privacy of k-mean clustering using homomorphic cryptosystem.
	Balebako et al. [18]	Detecting through filtering	Detects the leakage of privacy through filtering based on TaintDroid in the Android environment.
	Liu et al. [19]	Shadow coding schema	Data privacy through shadow matrix computation.
	Gupta et al. [20]	Abnormal detection	User classification through abnormal behavior detection and monitoring.
	Al-Shomrani et al. [21]	Sensitive data identification	Individual storage security module through sensitive data identification policy.
	Consolvo et al. [22]	Web traffic control	User access control by monitoring sensitive information related to keywords.
	Zhou et al. [23]	Access control	Privacy-preservation related to background application in android environments.
Tiwari et al. [24]	Detection data collecting	Data leakage detection and blocking through de-compile of APK files.	

3.1.1. Privacy Policy

In the case of securing data through active data collection, data must be collected in advance with the consent of the person generating the data. In addition, for a system that collects a large number of log records, the collection of data used internally is subject to the consent of the collection according to an internal policy for the subject of data ownership. The utilization of the data already possessed is also

a kind of data collection. However, at this time, consent to use the data already held must be obtained in advance, and the agreed data must be used for the purpose of use. In the case of passive data collection, data is generally collected through an automated system. It is necessary to obtain the consent of the data collected from the data owner and use the data, but it is difficult to obtain the consent of the data owner to collect separately because the collection process is automatic. After collection, it is possible to notice that the data is collected and used by the data owner. However, if the data to be collected is sensitive information including privacy, there may be legal issues, so the collection subject should be careful and collected according to the nature of the collected data. Data must be collected in the public domain, and if it is possible to infringe on the privacy of an individual by including sensitive information of the service user, data should be collected in consideration of legal matters.

GDPR is a legal regulation for data protection and privacy-preserving in the EU and the EEA. The GDPR aims to integrate EU regulations and control the access for personal data by simplifying the regulatory environment for international business. Data controllers and processors must take appropriate techniques to implement the data protection principles, and provide protection systems to guard data (e.g., pseudonymization and anonymization). The data controller should clearly disclose all collected data and state purpose in the data processing process. In addition, the data retention period, and data sharing with a third party must be specified, and the collected data must be able to be retrieved and destructed through the withdrawal of consent from the data subject. Greene et al. [11] indicated the need for the identification of GDPR concepts and principles, as the introduction of the GDPR in the EU makes it difficult for many companies and researchers to comply with regulations and collect data. They also describe how GDPR can impact the work of data scientists and researchers in this new data privacy regulation.

CCPA is general law for the state of California and is the most efficient privacy law in the US. Privacy defined by the CCPA is information that directly and indirectly identifies, related to a specific consumer or household. The major component is the obligation to notify when collecting, selling, and disclosing privacy. In addition, it describes the right to disclose, right to access, right to deletion, right to opt-out, and right to non-discrimination. Stallings et al. [12] described one aspect of the law that has received less attention in detail: how it deals with obfuscation algorithms that can protect privacy, including anonymization, aggregation, and pseudonymization. This paper considers whether these technical protections are well-defined enough and whether they are effective. Before considering technical protection in more detail, it is useful to first understand the scope of action. Three particularly important aspects are who the consumer is, which business applies, and what personal information means. Kanika et al. [13] suggested an extended data life cycle differently from previous studies, but there was no content on data destruction. However, according to laws dealing with privacy protection in various countries, such as the GDPR and the CCPA, when the information provider withdrawal of consent at any time or when the purpose of use is achieved. As it guarantees immediate destruction, the data destruction phase is a very important factor in the big data life cycle.

3.1.2. Privacy-Preserving Data Collection

Privacy-preserving data collection refers to a method that does not infringe on privacy when collecting various data. This includes security elements such as access control and homomorphic encryption.

Access control refers to the ability to allow or deny someone to use something. Determine the rights of users and services for files, printers, registry keys, and directory service objects. In other words, it is a method of guaranteeing that users accurately and reliably authenticate themselves and that they have the appropriate access to data. Access control includes authentication and authorization. Authentication is a technology used to confirm the identity information of an object that wants to access data, and authorization is a technology that grants access to an object based on authentication. In the collection phase, access control focuses on the privacy collected. Data collection can be provided directly or randomly based on the consent of the data provider. In the case of randomly collecting information, access control such as authentication and authorization for data access is required. Hornyack et al. [14]

proposed a system called AppFence that modifies the operating system in the Android environment to prevent indiscriminate data access by applications and protect privacy. It is a system that allows applications to block unnecessary access control requests and protect privacy by sending shadow data if the user does not want to. Zhao et al. [15] proposed a personal data cloud that collects and stores personal data of various services through Open API, web crawler, and manual importation. Open API method is divided into authorization-based and non-authorized-based, and authorization-based is a method of a limited collection with an access token through an authorization mechanism such as OAuth 2.0. In the case of data that requires authorization in the web crawler method, login emulation is required. At this time, it is said that access control can be realized through access control with other authentication and authorization mechanisms. Manual importation is a method in which the user directly provides data, and only the user can transfer the desired data.

Homomorphic encryption is an encryption method that can perform various operations without decrypting the ciphertext. It is the result of the operation is the same as the operation result of plain text, so the usefulness of data can be secured while protecting sensitive data of the data provider. Homomorphic encryption is generally used in the data collection phase rather than the analytics phase because the computational processing speed is very slow, and it is impossible to accurately decrypt in some cases. However, there is a lot of research going on, so it will be available in the analytics phase in a few years. Gao et al. [16] proposed a privacy-preserving auction scheme that enables data providers to sell data securely through OTP and homomorphic encryption and an enhanced privacy-preserving auction scheme with enhanced security. They have solved several problems such as manipulated messages and bad entities. It was a disadvantage of increasing processing time due to the added signature verification mechanism, but they suggested several ways to solve this issue. Mittal et al. [17] proposed an approach to mining while solving user privacy threats in the cloud environment. This approach protected privacy while maintaining the accuracy of k-mean clustering using a pallier homomorphic cryptosystem in a distributed environment. They discussed through security analysis that the proposed approach is safe against some attacks. However, this approach was difficult to apply in a centralized cloud environment.

In addition to access control and homomorphic encryption, various privacy-preserving data collection includes the following studies and methods. Balebako et al. [18] proposed a prototype that allows users to recognize privacy leakage based on TaintDroid in the Android environment. It detects the leakage of privacy through filtering, sends a notification to the user, and allows to check the information or number of data shared through the proposed application. Liu et al. [19] proposed a shadow coding scheme that collects data while achieving data privacy of distributed data providers. This is a way to protect the privacy of data through shadow matrix computation when collecting data and to recover data in case of failure. However, this method can only be used in a synchronous environment, and there is a limit to the convergence of various privacy protection requirements. Gupta et al. [20] proposed a model that detects abnormal or suspicious behavior through a library that detects abnormal behavior in the data collection phase, monitors the user's activity, blocks malicious users in advance, and distinguishes them from important users. However, there is a disadvantage that it is inefficient to check all the big data because the size of big data is exponential. In addition, frequent updates occur because values for abnormal behaviors must be updated in advance in the library. Al-Shomrani et al. [21] proposed storage where the information provider can identify sensitive data using a security module based on the policy by creating and sending policies that identify sensitive data when the information provider stores information in the cloud. There is a possibility of not being able to recognize sensitive data by itself, and there is a disadvantage in that a provider who is worried about leakage cannot provide a large number of information and thus cannot obtain highly useful mining results. Consolvo et al. [22] proposed a method to control unauthorized web traffic through a Wi-Fi privacy ticker. This is a method of performing monitoring based on the user's registration of sensitive information, and when such information is detected, the user is notified of this to perform access control. In addition, even if sensitive information is not registered, it can be detected when

personal data is transmitted without encryption. Zhou et al. [23] proposed taming information-stealing smartphone applications. The proposal is a privacy mode that can control application data access in the Android environment. This can achieve privacy by sending a request to allow the user to recognize the application requesting information without affecting the operation of the application and sending fake information upon rejection. In addition, this is a system that can precisely control access control for user information such as scope and method. Tiwari et al. [24] proposed a method to detect secret communication and prevent application data leakage in the Android environment. It can detect and block data leaks and secret notices after decompiling the APK file through reverse engineering.

3.2. Storage

The storage phase includes encryption, data partitioning, and access control. The related data storage studies are summarized in Table 4.

Table 4. Description of conceptual approaches for big data storage.

Type	Papers	Approaches	Descriptions
3.2.1 Encryption	Xu et al. [25]	ABE	Solve valid access after user revocation, exposure of temporary decryption key.
	Li et al. [26]	ABE	Creates an encrypted trapdoor for each keyword and decrypt without knowing the keyword.
	Xue et al. [27]	ABE	Complete deletions by using proxy re-encryption and Merkle hash tree.
	Yang et al. [28]	ABE	Data sharing between cross-domain and ensures that the same data can be safely deduplicated.
	Baek et al. [29]	IBE and proxy re-encryption	Replacing digital certificates with the identifier.
	Zhang et al. [30]	IBE	Prevents unauthorized access and periodically updates the secret key.
	Azougaghe et al. [31]	AES and ElGamal	Data encryption with AES and key encryption using ElGamal algorithms.
	Hussien et al. [32]	AES and ECC	Integrity verification through hash and data protection through ECC and AES.
	Li et al. [33]	SED2 algorithm	Divide data through intelligent encryption and make only it visible to the cloud provider.
	Al-Odat et al. [34]	Multi-authentication and SHA	Integrity guaranteed through SHA and multi-authentication method encrypted.
	Arora et al. [35]	Hybrid encryption	Combination of HMAC, OTP, SHA, symmetric key, and asymmetric key in the cloud environment.
	Saroj et al. [36]	Threshold encryption	Ensures confidentiality through threshold encryption and guarantees integrity.
	Bajwa et al. [37]	Obfuscation and encryption	Protects data by obfuscation and encryption according to the type of data.
	Sanka et al. [38]	Access control and encryption	Ensures only data owners and users can view the data through a symmetric key.
3.2.3 Access Control	Cheng et al. [39]	Paths encryption	Storage that encrypts the paths and protects data mapping through the trap door function.
	Al Hamid et al. [40]	Bilinear pairing cryptography	Secure communication and data protection using bilinear pairing cryptography.
	Saraiva et al. [41]	Evaluation of encryption algorithm	Presented an encryption benchmark to protect data among heterogeneous resources.
	Ko et al. [42]	Private cloud	Proposed a model to ensure confidentiality and privacy using a private cloud.
	Ngo et al. [43]	Distributed clouds	Role-based policy management using a policy profile in XACML and sharing security context.
	Yu et al. [44]	Attribute-based access control	Allows the data owner to hand over the work related to access control without providing data.
	Younis et al. [45]	Role-based access control	Allow secure data sharing and efficient access control through security tags and risk engines.
	Liu et al. [46]	Data access	Allows data users to access data containing specific keywords in the cloud through search.
3.2.4 Audit Trail	Adrienne et al. [47]	Data access	The privacy-by-proxy approach to achieve privacy.
	Sundaeswaran et al. [48]	Data logging	Method to create a JAR for the access policy.
	Yang et al. [49]	Various audit methods	Explained various audit methods and analyzed security and performance in detail.
Ferdous et al. [50]	Blockchain	Proposed an architecture that evaluates whether access control has been properly performed.	
Wang et al. [51]	Token-based method	Decentralized system method that allows data owners to detect data corruption through tokens.	

3.2.1. Encryption

Encryption is a method of transforming understandable data (e.g., plaintext) into incomprehensible form (e.g., ciphertext). This is to ensure that only authorized users can use the data. Technically, it refers to the process of converting plaintext into ciphertext through a mathematical algorithm. Only those who have a set of encryption keys can change the ciphertext to plaintext, which is called decryption. Encryption is still the most basic strong protection technology to ensure confidentiality. Representatively, it is divided into a public key algorithm and a symmetric key algorithm, and many studies use existing encryption methods for storing big data. This section describes representative ABE, IBE, RSA, and AES.

ABE is a type of public-key cryptography that performs encryption and decryption based on an object attribute set and the access structure. It is that decryption is possible only when the attribute of the ciphertext and the user attribute set match. ABE is divided into KP-ABE and CP-ABE. In KP-ABE, the conditions (e.g., policy) that can be decrypted are included in the user secret key, and CP-ABE is included in the ciphertext; it is vulnerable to collusion attacks. In addition, it is widely used in IoT environments with many elements that can be used as attributes. Xu et al. [25] suggested CP-ABE, pointing out the limitations of existing attribute-based encryption in the IoT Cloud. It solved the problems that occur in IoT cloud environment such as valid access after user revocation, exposure of temporary decryption key. Li et al. [26] proposed a keyword search function outsourced ABE that can solve shortcomings such as inefficient query processing when using ABE in a cloud environment. The proposed method creates an encrypted trapdoor for each keyword, the cloud service provider was able to search and partially decrypt without knowing the keyword and plain text. In addition, it saved a chosen-plaintext attack and has scalability and efficiency. Xue et al. [27] proposed a key policy ABE scheme for secure deletion to solve the problem of data not being completely deleted in a cloud environment. The algorithm made possible complete deletions by using PRE and Merkle hash tree algorithm. They also discussed the algorithm's security through an attribute-based selective-set model. Yang et al. [28] proposed a storage system that preserves the privacy of healthcare big data and enabled flexible access control. The system enabled data sharing between cross-domain using ABE and ensured that the same data can be safely deduplicated. It was designed to recover past healthcare data in emergencies using the break-glass access method.

IBE is a type of public-key encryption first proposed by Adi Shamir. IBE method generates a public key for the user identity (e.g., email and phone number). This allows secure communication without additional authentication. It is also efficient as it does not require certificate management. However, since the private key is generated through the key generation center, which is a third-party server, there are several problems such as reliability. Baek et al. [29] proposed a cloud-based framework, smart-frame, for managing big data in a smart grid. The framework consisted of three hierarchical structures and had high scalability. In addition, IBE, PRE, and signature are used to secure communication. They solved the issues by replacing digital certificates with identity to solve the digital certificate management problem, which is the drawback of IBE. Zhang et al. [30] proposed a cloud-based secure big data storage system. The system prevented unauthorized access through continuous leakage resilient IBE and ensured the confidentiality of data even in case of partial secret key leakage. It was achieved by periodically updating the secret key in the proposed system. They have proven safe in the continuous leakage model as they have a high leakage ratio of 1/3 against leakage attacks.

RSA is a well-known public-key cryptography, made by three mathematicians (e.g., Ron Rivest, Adi Shamir, and Leonard Adleman). RSA utilizes the difficulty of prime factorization in encryption and decryption. Public-key cryptography has two secret keys (i.e., the public key and a private key). Anyone with a public key can encrypt data, but it can only be decrypted using a private key. Due to its low speed, it is not used to directly encrypt sensitive data but is generally used to transmit the secret key of symmetric key cryptography. It can be used in a digital signature form to prove the authenticity and integrity of a message. Currently, it uses in secure socket layer/transport layer security.

AES is a symmetric block encryption algorithm adopted by NIST in the US. This is an encryption method based on the Rijndael algorithm selected through a public offering. AES consists of a substitution permutation network and has stability in linear cryptanalysis and differential cryptanalysis. It is suitable for both hardware and software encryption with sensitive data. By default, the length of the encryption key can be extended to 128 bits, 192 bits, or 256 bits. Due to its fast encryption and decryption speed and stability, it is widely used for large data storage and database encryption to this day. Azougaghe et al. [31] proposed a simple cloud storage protection method that encrypts data to be stored in the cloud through AES and encrypts. In addition, the key encrypts using the ElGamal algorithm and stores it in the intern server. Hussien et al. [32] proposed cloud storage that can guarantee data integrity and security in a cloud environment through AES, Hash Algorithm, and ECC. This enables a third-party auditor to avoid untrusted CSPs through data integrity verification and hashing on behalf of data owners. To respond flexibly to man-in-the-middle attacks.

In addition to ABE, IBE, RSA, and AES, various encryption includes the following studies and methods. Li et al. [33] proposed security-aware efficient distributed storage, in which cloud service providers do not have direct access to sensitive data in the cloud. Among the three algorithms used, the first alternative data distribution is an algorithm that determines whether data packets should be divided and stored in a distributed cloud server to shorten the operation time, and the second secure efficient data distributions is an algorithm that performs intelligent encryption before being stored in the cloud. It is an algorithm that divides data into two and stores only one information in storage that the cloud service provider can see and the other is not visible to anyone. Finally, efficient data conflation is an algorithm that allows you to obtain the desired information by recombining divided information. Al-Odat et al. [34] proposed a secure distributed big data storage through Shamir's algorithm and SHA. This makes it impossible to decrypt data even if one part of the encryption key is obtained through Shamir's secret sharing, enables multi-authentication, and guarantees data integrity using SHA. Arora et al. [35] proposed a hybrid encryption system with a strong encryption process that can be used in a cloud environment using various encryption technologies such as HMAC, OTP, SHA, salting, symmetric key algorithm, and asymmetric key algorithm. However, since it uses various encryption techniques, it may reduce the efficiency of the system, and has a disadvantage that it cannot be used in a multi-cloud environment. Saroj et al. [36] proposed a method for data owners to protect the confidentiality of data in the cloud by using encryption based on threshold encryption in a cloud environment. This ensures the confidentiality of the data through threshold encryption and guarantees the integrity of the data by encapsulating the message digest along with the data in the encryption process. In addition, a nonce based on the Diffie–Hellman key exchange algorithm and a key exchange algorithm using a session key was used to ensure confidentiality in the transmission process. Finally, a capability list was created for data access control, and a method of access control only by authorized users was proposed. Bajwa et al. [37] proposed a data owner-centered cloud data protection method. It protects data by obfuscation and encryption according to the type of data and provides role-based access control. In addition, the integrity of the message can be guaranteed through HMAC. Sanka et al. [38] proposed an access control and encryption method to protect data in a cloud environment. This ensures confidentiality so that only data owners and users can view the data through a symmetric key encryption algorithm, and the integrity of the data through a hash function. In addition, additional encryption is performed through the Diffie–Hellman key exchange algorithm using a shared one-time session key to protect data during transmission. However, this may increase the maintenance and security problems of the symmetric key. Cheng et al. [39] proposed storage that can be shared with other users for data paths, which can store data by dividing it into sequence parts through a distributed cloud environment, encrypt the paths and protect data mapping through the trap door function. However, since this only protects the path of the data, there is a possibility that data may be damaged due to malicious data modification. Al Hamid et al. [40] proposed a method to secure medical data in the fog cloud environment. The proposed method used bilinear pairing cryptography to achieve secure communication and data confidentiality protection between participants. They used

honeypot and decoy techniques to confuse the attacker. Saraiva et al. [41] presented an encryption benchmark to protect data among heterogeneous resources. As encryption methods, various symmetric key algorithms were compared and evaluated using all key sizes and authenticated encryption modes. They compared and evaluated encryption algorithms by measuring execution time, throughput, battery consumption, and security vulnerability. Based on it, they evaluated the preference and reliability of the optimized latest encryption algorithms.

3.2.2. Access Control

In the storage phase, it is divided into physical access control to storage and logical access control to stored data. Physical access control is required to prevent attacker access and extortion to the storage. In addition, access control techniques and policies are required so that only users who are authenticated and have the authority to access the stored data can access them. Ko et al. [42] proposed a hyper execution model that protects confidentiality and privacy in the cloud. The model classifies the sensitivity of the data prior to calculation, calculates non-sensitive data in the public cloud, and calculates the organization's sensitive data (e.g., personal data) using the private cloud. Therefore, it provides integration with safety. Ngo et al. [43] proposed a method of constructing a security infrastructure that supports consistent trust establishment, access control, and context security management by following a general service life cycle management model in a virtual cloud environment that provides infrastructure as a service. This enables role-based policy management using a policy profile in extensible access control markup language, and it can solve the problem of sharing security context between distributed clouds through an authorization ticket technique. Yu et al. [44] proposed a method for attribute-based access control in a cloud environment. This, by combining PRE with KP-ABE, allows the data owner to hand over the work related to access control to the cloud without providing data information, thereby reducing the overhead of the data owner and enabling easy access control. The computational overhead of the cloud can be reduced by aggregating computational tasks through encryption. Younis et al. [45] identified access control requirements (i.e., trust and scalability) in a cloud environment and discussed the limitations of traditional access control models such as mandatory-based access control. In addition, they proposed a new model that meets the requirements. The proposed model allowed secure data sharing and efficient access control through security tags and risk engines. Liu et al. [46] proposed secure and privacy-preserving keyword searching that can be used in a cloud environment. This is a method that allows data users to access data containing specific keywords in the cloud through search, regardless of where they are using any device. CSP is involved in decryption to reduce the load on the data searcher and check the data before returning results. The accuracy of the data can be guaranteed. Adrienne et al. [47] proposed the privacy-by-proxy approach to achieve privacy when making their data accessible to third parties in a social networking environment. This can be provided by abstracting data through special markup tags. When providing data, access control can be achieved through authorization checks in the proxy server. Sundareswaran et al. [48] proposed a method for data logging as well as data access control through JAR in a cloud environment. This is a method for the data owner to create a JAR for the access policy for the data to be published, digitally sign it, and hand it over to the cloud, and then the cloud uses it to control access. The advantages of this study are that when data is accessed, the user's digital signature enters the logging file, increasing the reliability of logging, enabling backtracking in case of a problem, and maintaining strong backend security upon request by the owner. It has the advantage of creating copy files and verifying data integrity.

3.2.3. Audit Trail

Audit Trail is essential at all phases in the big data life cycle. In particular, it is even more important in the storage phase because it can be the case that sensitive data from the data provider has been saved. Therefore, storage administrators need to log who, when, where, and what requests were made on the storage, and a rigorous audit of whether the system responded appropriately. In addition, because big

data is mostly stored in a distributed environment, it is important to audit data storage and the integrity of the data. Yang et al. [49] analyzed various storage auditing services in a cloud environment. Especially, they explained the need for a third-party auditor because increases the possibility that a fair result may not be possible if the storage administrator and data provider are asked to conduct an audit. The detailed audit methods (i.e., MAC-based methods, RSA-based homomorphic hash value) and analyzed security performance. Ferdous et al. [50] proposed a blockchain-based decentralized runtime monitoring architecture for a distributed access control system. It is an architecture that evaluates whether access control has been properly performed according to the policy being used and detects policy violations by storing logs and monitoring based on the blockchain. However, this has a disadvantage in that it may take a long time for smart contracts and monitoring due to the size of the log. Wang et al. [51] proposed a decentralized system method that ensures data integrity even in a distributed cloud environment, allows data owners to detect data corruption and malfunctioning nodes through tokens, and guarantees data reliability even in various malicious attacks.

3.3. Analytics

The analytics phase includes privacy-preserving data mining and access control. The related studies of data analytics are summarized in Table 5.

Table 5. Description of conceptual approaches for big data analytics.

Type	Papers	Approaches	Descriptions
3.3.1 Privacy preserving data mining	Mohan et al. [52]	Association rule hiding	Proposed hiding techniques based on genetic algorithm and dummy items creation technique.
	Gopalan et al. [53]	Association rule hiding	Developed an efficient meta-heuristic algorithm based on the chemical reaction optimization algorithm.
	Menga et al. [54]	Association rule hiding	Proposed secret key generation method using the least lion optimization algorithm.
	Liu et al. [55]	Aggregation	Proposed a practical privacy-preserving data aggregation scheme without TTP.
	Gahar et al. [56]	Reduction	Reduction algorithm based on the MapReduce paradigm.
	Motiwalla et al. [57]	Data masking	Protects privacy without removing the attributes of the data and delivers it to necessary third parties.
	Cui et al. [58]	Data masking	Masking method based on format-preserving encryption in a distributed environment.
	Geo et al. [59]	Differential Privacy	Protects privacy through differential privacy when performing the k-means clustering process.
	Ni et al. [60]	Differential Privacy	Initial objects are randomly selected, and privacy protection can be realized using Laplace noise.
	Mo et al. [61]	Differential Privacy	A data preprocessing method based on differential privacy for distance-based clustering.
	Zhao et al. [62]	Differential Privacy	A privacy protection method that can be used in distributed collaborative mining.
	Lin et al. [63]	Differential Privacy	Concept of a dynamic noise threshold to analyze the relationship between the noise size and the data set.
	Zhang et al. [64]	k-anonymity	MapReduce parallel processing to perform anonymizing through k-anonymity.
	Mehta et al. [65]	k-anonymity	Protect big data publishing without a specific mapper and reducer.
	Machanavajjhala et al. [66]	l-diversity	Proposed a novel and powerful privacy definition called l-diversity.
	Li et al. [67]	t-closeness	Protects privacy by making the difference between the distribution of sensitive information.
	Vatsalan et al. [68]	Privacy-Preserving Record Linkage	Presented a survey of privacy-preserving record linkage technologies in the past and present.
Scannapieco et al. [69]	Record Matching Protocol	Presented the protocol that allows each object to hide records, schema attribute details that are not shared.	
Gkoulalas-Divanis et al. [70]	Border-based Approach	Applied that to hide sensitive data sets and introduced minimal extensions to the original database.	

3.3.1. Privacy-Preserving Data Mining

PPDM method refers to a technology that finds knowledge or patterns implicit in data without infringing on the privacy of data owners. There are two types of PPDM: the method of analyzing by adding noise to the original data or applying randomization, and the SMC method, in which information other than input and calculation results cannot be obtained. The anonymization analysis method has been practically used for various statistical data but has the disadvantage of being vulnerable to security. In addition, due to low calculation efficiency, SMC is not practical. Therefore, the PPDM method needs to be selected continuously due to the trade-off of the safety and practicality of calculation. PPDM includes statistical disclosure limitation, association rule hiding, homomorphic encryption, de-identification, and privacy models.

Association rule hiding is an algorithm to prevent creating sensitive association rules in the analytics phase. Association rules have a high usability in various mining algorithms. However, it can identify individual sensitive data through the association rule. Therefore, association rule hiding prevents sensitive association rules from appearing with minimal modification of sensitive data, in ways such as deleting and adding data values. Generally, association rule hiding algorithms are classified as heuristic approach, border-based approach, etc. [71,72]. Mohan et al. [52] proposed a genetic algorithm-based hiding technique and a dummy item creation technique. They aimed to prevent the identification of sensitive association rules by creating a genetic algorithm-based hiding technique and a dummy item of modified sensitive information. In particular, they modified sensitive items to protect the linking rules and used dummy item creation to keep the same cost for the original and new databases. Gopalan et al. [53] developed an efficient meta-heuristic algorithm for association rule hiding based on a chemical reaction optimization algorithm. The results of the proposed approach are compared with the genetic algorithm, particle swarm optimization, and cuckoo-based algorithms. The experimental results of the proposed algorithm are tested on the benchmark datasets. Menaga et al. [54] proposed a technique of secret key generation for privacy-preserving using the LLOA. The proposed algorithm involves rule mining and secret key generation for the sanitization. Initially, the whale optimization algorithm mines the association rules for the input database and validates the rules with the newly formulated fitness function. An algorithm, LLOA is developed by modifying the lion optimization algorithm with the inclusion of the least mean square which generates a secret key to provide privacy in mining. With the secret key, LLOA converts the original database into the sanitized database. Then, the algorithm optimally selects a secret key such that the sanitized database hides sensitive information by the utilization of two factors, namely, privacy factor and utility factor, in its objective function.

De-Identification is a method of deleting the PII in data or replacing it with attribute information. The major purpose of de-identification is to ensure that data including privacy can be combined with other data so that a specific individual cannot be identified. De-identification should be applied at all phases of the big data life cycle, such as collecting, storing, utilization, and sharing of privacy. Various methods and algorithms are included in the de-identification process. In this paper, data masking and data filtering are described as de-identification methods when storing data.

- *Pseudonymization* refers to processing so that a specific individual cannot be recognized without additional information by deleting part of privacy or replacing the part. When processing a pseudonymization, it is necessary to consider whether a specific individual can be recognized by the pseudonym information and the possibility of combining additional information. Typical pseudonymization techniques include encryption, hashing, and tokenization.
- *Aggregation* is a de-identification technique of making the values of a sensitive data set into average or total values to prevent the identification of sensitive data values. When used in the analytics phase, the usefulness of the data is reduced, and detailed analytics is difficult. Therefore, it is necessary to collect a lot of data to ensure the accuracy of mining. Generally, aggregation uses methods such as micro-aggregation, rearrangement, and rounding. Liu et al. [55] proposed a privacy-preserving data aggregation method that does not rely on a TTP. Most existing data

aggregation methods rely on TTP and have security issues such as a denial of service attacks. They described the data aggregation model in the smart grid domain and configured the data collection unit to form a virtual aggregation area. The aggregate result is masked and used for data analysis. In addition, by reducing the aggregation area, some defects of the aggregation operator are negligible. They focused on developing solutions that balance data utility and privacy, ensuring that the aggregate results have little impact on data utility.

- *Data reduction* is a direct method of erasing sensitive data. Generally, values that can directly identify the data provider such as zip-code, e-mail, and social security number, are temporarily or completely deleted to make them unidentifiable. The data reduction is not used much other than direct identifying information such as PII. Gahar et al. [56] focused on the performance degradation due to missing data of the existing statistical algorithms. To solve the data missing problem, they proposed a reduction algorithm based on the MapReduce paradigm of the RHadoop framework. They approached a distributed statistic method and used a random forest imputation method. The proposed algorithm is based on PCA and MCA. The PCA method processes quantitative variables and the MCA method processes categorical variables. In addition, it facilitates data search by reducing the search space in the process of extracting useful information.
- *Data suppression* is the conversion of data values into grouped values. For example, if the value is 35, it is converted to a value of 30–40. This makes it difficult to ensure accurate mining results with larger grouping ranges. However, data suppression difficult generally infers the original value of the data set and does not have a huge impact on data usability.
- *Data masking* is the most actively used method of de-identification. This is usually de-identification by combining sensitive data from the data provider with other data or replacing parts of the data. There are various techniques such as substitution, shuffling, and nulling. There are two main types of data masking such as static data masking and dynamic data masking. Motiwalla et al. [57] proposed a system that protects privacy without removing the attributes of the data through a masking technology for healthcare data and then delivers it to necessary third parties. Cui et al. [58] proposed a method of masking data while maintaining the format of sensitive data (e.g., date and e-mail) based on FPE in a big data environment. This method can be applied in both single and distributed environments. In particular, it can achieve high efficiency in a distributed environment. However, compared to the symmetric-key algorithm, the speed was significantly slower. In addition, this cannot be preserved the association of data.

Privacy model is a model to prevent the data subject from being identified by de-identifying information (e.g., PII, zip-code, and birthdate) that may identify the subject of the information in case of attempting to disclose the datasets to a third party for public purposes such as research. The model includes k-anonymity, l-diversity, and t-closeness, and the main purpose is to achieve public interest by limiting the level of identification threat caused by leakage.

- *Differential privacy* is a mathematical model for preventing privacy inference based on query results performed in a statistical database, and various related studies are being conducted to protect the privacy of statistical data. This method is one of the PPDM that maintains the distribution of data and adds noise without harming the original statistical meaning. The information exposure is limited by keeping the amount of change in query results according to insertion, deletion, and transformation of data below a certain level. If the query result changes significantly due to the change of the information of a specific individual, the attacker can see the difference in the query result and know the existence of data of a specific user and the value of the data. Differential privacy applies to online inquiry systems, and it is also possible to use differential privacy to generate machine learning statistical classifiers and synthetic information. Gao et al. [59] proposed differential privacy hybrid k-means that protects privacy through differential privacy when performing the k-means clustering process in the Apache Spark environment. This improves k-means clustering by combining the swarm intelligence optimization model and additionally

protects privacy through the Laplace mechanism that adds noise. Ni et al. [60] proposed a schema for privacy-preserving by using differential privacy in the process of PPDM through DBSCAN. This is a method of performing mining by determining several core objects, unlike DBSCAN, in which initial core objects are randomly selected, and privacy protection can be realized through noise technique. Mo et al. [61] proposed a data preprocessing method based on differential privacy for distance-based clustering. The adaptive parameter mechanism used here is a preprocessing method that maintains a balance between privacy protection and clustering results. This is a mechanism in which the higher the security function, the higher the privacy protection strength, and the higher the availability of data function, the higher the availability of data. Zhao et al. [62] proposed a privacy protection method that can be used in distributed collaborative mining. This allows individual data owners to protect privacy by using differential privacy in the regression and classification learning process, and to ensemble the information of various trees through a gradient boosting decision tree while protecting privacy without third parties. Lin et al. [63] focused on the problem of exposing sensitive information in the existing big data collection method and proposed a differential privacy protection system in a body sensor network environment. The proposed model introduces the concept of a dynamic noise threshold to analyze the relationship between the noise size of electrocardiogram data and the size of the data set. As the proposed model can perform sufficient interference with the data, even if the attacker fully knows the background, it cannot find a match with a specific victim.

- *k-anonymity* is one of the privacy-preserving models to prevent linkage attacks by linking public information and is used to prevent re-identification of de-identified privacy. *k-anonymity* refers to a measure that, when de-identifying, ensures that there is at least k or more of the same value in a given data set so that they cannot be easily combined into other information. It is particularly effective in protecting the privacy of data with limited properties and hides sensitive data with generalization, containment, analysis, and permutation techniques. However, when de-identifying, the diversity of information is not considered. When records with the same information are de-identified and composed into a single set, there is a limit that is defenseless against homogeneity attacks. Zhang et al. [64] proposed a two-phase top-down specialization approach that anonymizes big data using MapReduce in a cloud environment. This is a method of dividing big data into small data using MapReduce parallel processing to perform primary anonymization, and then anonymizing it once more through *k-anonymity*. However, there is a possibility that processing efficiency is difficult in that a lot of data is anonymized twice. Mehta et al. [65] proposed a MapReduce-based scalable *k-anonymization* algorithm. The major purpose of the proposed algorithm is to simplify the approach using Apache Pig and to protect big data publishing without specific mapper and reducer programs. In addition, it divides the data set into smaller than existing algorithms based on all the attributes of the data set, and utilizes sorting and shuffling for data distribution and merging in Hadoop. Therefore, it reduces the number of iterations compared to the existing algorithm, shortens running time, and performs the same level of privacy-preserving.
- *l-diversity* is a model for defending against homogeneity attacks against *k-anonymity*. Even if *k-anonymity* is satisfied, a small number of categories increases the likelihood of being identifiable. *l-diversity* means that records that are de-identified in a given data set must have at least l different sensitive information. Even if it is de-identified by the *l-diversity* model, *t-closeness* is required to prevent skewness attack and similarity attack. Machanavajjhala et al. [66] suggested that if various attributes do not exist in sensitive data and the attacker has background knowledge, *k-anonymity* cannot guarantee disclosure of sensitive information about the attacker. Therefore, they described the possibility of an attack in both cases and proposed *l-diversity* to complement the problem. *l-diversity* means that records that are de-identified in a given data set must have at least l different sensitive information.
- *t-closeness* is a model to overcome the weakness of *l-diversity*. If the information in the records is skewed or similar to each other, there is a problem that privacy is exposed through the difference

in the distribution of sensitive information. Therefore, t -closeness is a model that protects privacy by making the difference between the distribution of sensitive information of records that are not identified from the data set and the distribution of sensitive information of the entire data less than t . The closer the t value is to 0, the stronger the similarity between the distribution of the entire data and the distribution of a specific data section tends to be stronger. That is, the distribution of specific information in each homogeneous set is not too specific compared to the distribution of the entire data set. Li et al. [67] described the limitations of l -diversity and proposed a new privacy concept to overcome them. The l -diversity can allow an attacker to disclose privacy when information is skewed to a specific value and when de-identified information is similar to each other. To solve the problem of l -diversity, they proposed the concept of t -closeness. t -closeness is a model that protects privacy by making the difference between the distribution of sensitive information in an unidentified record in a data set and the distribution of sensitive information in the entire data by less than t .

In addition to association rule hiding, de-identification, and privacy model, various PPDM methods include the following studies and methods. Vatsalan et al. [68] provided an overview of privacy-preserving record linkage, a technology that allows database connections between organizations while protecting data privacy. They presented a survey of state-of-the-art technologies in the past and present on this technology. In addition, they identified the classification of the technology and identified it in 15 dimensions. Through this classification, they identified various shortcomings of the current approach. Scannapieco et al. [69] proposed a record matching protocol that protects privacy at the data and schema level. In particular, if different objects need to identify common data, the proposed protocol allows calculating the match of a data set without sharing exact data. This protocol allows each object to hide records, schema attribute details that are not shared. Gkoulalas-Divanis et al. [70] proposed a border-based approach to provide a solution that can hide sensitive and frequently used data sets. They utilized cover relationships between revised borders and data sets to keep sensitive knowledge. In addition, they concealed and minimized data sets included in the constraint satisfaction problem. Finally, they applied binary integer programming to hide sensitive data sets and introduced minimal extensions to the original database.

3.3.2. Access Control

In the analytics phase, the data provider's sensitive data may be infringed by the data analyst. Therefore, it is necessary to ensure that the analytics is performed by a data analyst, who is certified and has valid authorization. In addition, appropriate access control policies and techniques must be implemented to prevent out-of-purpose analytics.

3.4. Utilization

The utilization phase includes audit trail and privacy-preserving data publishing, and Table 6 summarizes related studies on data utilization.

3.4.1. Audit Trail

In the utilization phase, various privacy issues can arise when the results derived from the previous phase (i.e., analytics phase) are disclosed to the public (e.g., disclosure to researchers) or used for achievement of purpose. Therefore, it is necessary to record who uses the data, how and where it is appropriately used. In addition, when an auditor wants to know why a certain decision was made using a machine learning model, an audit trail is used to trace back to the factors on which the model decision was based.

3.4.2. Privacy-Preserving Data Publishing

One of the models for using privacy in a database while protecting the privacy of the information subject is PPDP. The purpose of using PPDP is to provide new unidentified or synthesized information

that can be distributed to users without exposing the identity of the data subject. In other words, PPDP can be used to publish information based on privacy, allowing other researchers to conduct new analyses. PPDP includes de-identification and visualization techniques. Dasgupta et al. [73] proposed privacy-preserving visualization in parallel coordinates. The model used the choice of a distance metric and locality-preserving clustering as the clustering algorithm and k-anonymity and l-diversity to preserving privacy. They restricted direct access of users to data through an interactive interface and provided visualization tools. Finally, they discussed the potential attack and threat scenarios. Dasgupta et al. [74] have laid the research foundation for privacy-preserving visualization by identifying privacy threats and attacks that can occur in various visualization methods used in the visualization of electronic health data. They presented open challenges, such as the proper combination of existing privacy-preserving technologies and visualization. Chou et al. [75] discussed two privacy threats and proposed an interface that can achieve privacy protection in the visualization process based on the privacy model (i.e., k-anonymity, l-diversity, and t-closeness). This interface is designed not only to identify potential privacy breaches but to balance the utility and privacy of data at user request.

Table 6. Description of conceptual approaches for big data utilization.

Type	Papers	Approaches	Descriptions
3.3.1 Privacy-preserving data publishing	Dasgupta et al. [73]	k-anonymity and l-diversity	Protect privacy in visualizations using parallel coordinates. The method provided an interactive interface that could prevent direct data access.
	Dasgupta et al. [74]	privacy-preserving visualization	Discussed privacy-preserving visualization in health data, pointing out the limitations of several studies.
	Chou et al. [75]	k-anonymity, l-diversity, and t-closeness	Interface for privacy-preserving visualization that can detect potential privacy issues and increase data utility.

3.5. Destruction

In the data destruction phase, data used for data analysis is deleted. Data such as privacy must be destroyed without delay after the purpose of use is achieved, unless otherwise specified in other laws. There are data destruction solutions that involve the physical destruction of the hard disk and erasure by overwriting data in the existing database, based on degaussing and overwriting. Since most of the studies do not describe techniques for data destruction, this section emphasizes the necessity of this technology by taking examples related to data destruction. Toysmart is an electronic retailer that sold educational children's toys and has posted a privacy policy stating that information collected from customers is not shared with third parties. However, with a filing for bankruptcy in 2000, an attempt was made to sell customer information to a third party that should have been destroyed appropriately. Accordingly, a lawsuit was filed, and the federal trade commission ruled that opt-in consent was required to change the purpose of use of the collected information and that all customer information should have been deleted and destroyed, if it were impossible to delete [76]. In the case of privacy protection and electronic documents, a bank user requested the deletion of the social insurance number, and the bank communicated in writing that the user information was deleted. However, the information was not deleted, and the bank user filed a lawsuit. Accordingly, the commissioner stated that the bank violated principles 4.3 and 4.3.8 [77].

4. Evaluation

As mentioned previously, we proposed the taxonomy of security and privacy in the big data life cycle in Section 3 and listed related techniques of the components. To evaluate our proposal, we collected survey studies conducting technical research for security and privacy-preserving in the big data life cycle. This section compares our proposed taxonomy with the survey studies that addressed security and privacy-preserving in the big data life cycle. We evaluated that the components of each phase of the big data life cycle are covered in each study, as shown in Table 7. All related

studies investigated and analyzed the techniques from the collection to utilization phase of the big data life cycle. The studies that addressed the privacy policy of the data collection phase are [78,79], and there are no studies that addressed privacy-preserving data collection related technologies. In data storage phase, encryption is addressed in all studies excluding [80–83], and access control in storage phase is addressed in all studies excluding [79,80,84]. In addition, the studies addressed audit trail are [78,82,85–87]. In the data analytics phase, technologies related to privacy-preserving data mining are investigated in all studies excluding [87,88], and the studies addressed access control is [78]. The studies addressed the audit trail in the data utilization phase are [78,84], and the studies addressed privacy-preserving data publishing-related technologies are [80,81,84,86,87]. Techniques related to data destruction such as degaussing, and overwriting are not covered in all survey studies. In most studies, data collection and destruction studies were performed less than in the storage and analysis. They assumed the reliability of data sources in the data collection phase and focused on the storage and analytics phase. However, when collecting data, verifying the reliability of the source and confirming the privacy protection law is an issue that should be considered the most important because it affects the phases of storing, analyzing, utilization, and destruction. In addition, although data destruction related standards and laws exist, it is difficult to apply existing technologies due to the characteristics of big data (i.e., distributed storage and web storage such as the cloud), and the development of technologies are insufficient. Therefore, the destruction phase and related technologies are not covered in most of the studies. We address these security issues and open challenges in the following Section 5.

In addition, we describe the contributions and summaries of each study as follows. Fang et al. [89] analyzed the latest development of privacy-preserving technology based on big data applications. They also classified the characteristics of privacy-preserving and basic conceptions, metrics, and the research direction of privacy. Some special aspects of privacy-preserving, such as access control, encryption, anonymous protection, data auditing, and differential privacy protection, are discussed in detail. Finally, they introduced the privacy-preserving problem with its social implications and described some open challenges. Ye et al. [80] explained the big data characteristics: volume, diversity, speed, value, and authenticity, and categorized privacy and security issues into infrastructure security, data privacy, and data management. They detailed the privacy trajectory posting techniques (e.g., generalization, differential privacy, and deterrence) and discussed related research. However, they did not map security and privacy issues to the life cycle. Xu et al. [84] identified privacy issues from the data mining perspective and described techniques for various PPDM. They identified users in data mining applications as data providers, data collectors, data miners, and decision-makers, and discussed the privacy issues that may arise from each user. Finally, they explained the attack model related to privacy and discussed the mechanism for privacy protection and game theory. Jain et al. [85] presented the big data life cycle (i.e., data generation, data storage, and data processing) and identifying the privacy requirements in each phase. In addition, it explained the de-identification method (e.g., privacy-preserving aggregation and k-anonymity) to solve the requirements. They explained the fast anonymization of the big data stream environment, the recently used big data privacy protection technology (e.g., differential privacy and identity-based anonymization) explained in detail. Finally, they discussed the open challenges that may arise in fields that use big data such as healthcare and IoT. Abouelmehdi et al. [78] analyzed the security and privacy issues of big data in the medical field and discussed solutions. They presented the big data life cycle in the medical field (i.e., data collection, data transformation, data modeling, and knowledge creation) and explained in detail the technologies used for security measures in each phase, such as data masking, encryption, and access control. In addition, explained the privacy-preserving method, and discussed the characteristics of each country's laws for protecting medical data. Finally, they analyzed the limitations of several proposed studies to suggest future research directions. Yu et al. [79] discussed the various privacy-preserving technologies while explaining the difficulties of protecting privacy in the era of big data. They categorized privacy into content privacy and interaction privacy and discussed the 3 phases of work (i.e., collecting, anonymizing, and communicating) in the privacy

system. They classified privacy studies into data clustering (e.g., k-anonymity and l-diversity) and theoretical framework (e.g., differential privacy and membership privacy). In addition, they explicated privacy in terms of mathematics. Despite privacy is heavily influenced by the law, they have not been explained in detail. Jiang et al. [90] discussed big data applications and core technologies in a smart grid environment. They identified big data security requirements in a smart grid environment as privacy, integrity, authentication, and third-party protection, and discussed various solutions through related studies. Finally, they presented the open challenges of energy big data (e.g., data uncertainty, uncertain data mining, quantum cryptography, and data querying). However, they focused on the big data analytics phase. Mehmood et al. [91] presented the big data life cycle (i.e., data generation, data storage, and data processing) discussing the privacy mechanisms and infrastructure available in each phase. In particular, they elaborated on various privacy-preserving mechanisms such as access control, encryption, suppression, privacy-preserving clustering. Finally, they have presented some open challenges related to big data privacy. Alshboul et al. [92] proposed the big data life cycle (i.e., data collection, data storage, data analytics, and knowledge creation), discussed possible security threats and attacks at each phase such as re-identification, phishing, and spoofing. They also suggested that each phase should be countered to these threats through encryption and access control. Finally, they presented open challenges that effective security measures are needed because there is the potential to extract sensitive data through data mining. However, they did not elaborate on threats and security technologies. Moreno et al. [93] described big data and presented four big data security challenges (e.g., infrastructure security, data management, data privacy, integrity, and post-security). In addition, for each challenge, they identified the main topics of interest to researchers, such as authentication, access control, anonymization, laws, and cryptography. They remarked about the big data life cycle, but they do not make it clear. Wang et al. [81] explained the difference between privacy and security and the necessity of privacy-preserving in big data. In addition, while discussed privacy attacks such as correlation attack, differencing attack, reconstruction attack, and linking attack. They discussed models and mechanisms that protect privacy such as suppression, k-anonymity, and swapping. Especially, they explained differential privacy and discussed some of the limitations. they presented a research direction for privacy-preserving from the communication perspective. However, they lack an explanation of privacy protection technology used in the collection and utilization phases. Lv et al. [86] proposed the big data life cycle (i.e., data collection, data storage, and data application) and discussed the big data model architecture. They also identified the challenges at each phase and their security and privacy requirements (e.g., confidentiality and authenticity). Finally, they proposed some considerations (e.g., scalability, practicality, and balance) to solve big data security and privacy issues. Even though they presented a big data lifecycle, they did not map the protection technology to the big data life cycle. Goswami et al. [87] discussed the big data processing framework, security, and privacy challenges. They divided PPDP into two phases (i.e., data collection and data publish) and explained security techniques such as role-based access control. In addition, they explained data anonymization (e.g., suppression, k-anonymity, and t-closeness). However, they did not elaborate on the technology for privacy and security. Sangeetha et al. [82] discussed various big data platforms and big data architecture using them. the techniques for PPDM are classified and explained into input privacy (e.g., k-anonymity, differential privacy) and output privacy (e.g., association rule hiding and classification accuracy). In addition, they analyzed in detail the research to privacy-preserving of big data. Bertino et al. [88] identified big data requirements by some of the characteristics of big data. They discussed security techniques for confidentiality and privacy that meet security requirements. Especially, they discussed in detail the privacy and security issues that may arise in scenarios for IoT and online social networks. Finally, they mentioned the big data life cycle's importance but not presented the big data life cycle. Alwan et al. [83] presented a big data life cycle (i.e., data collection, data cleaning, data classification, data modeling, and data delivery). In addition, they analyzed big data in specific domains such as smart grid and IoT. Finally, they explained big data security. However, they did not explain it in detail and did not present privacy issues.

Table 7. Comparison of our proposal and related survey studies.

Life Cycle	Type	Fang et al. [89]	Ye et al. [80]	Xu et al. [84]	Jain et al. [85]	Abouelmehdi et al. [78]	Yu et al. [79]	Jiang et al. [90]	Mehmood et al. [91]	Alshboul et al. [92]	Moreno et al. [93]	Wang et al. [81]	Lv et al. [86]	Goswami et al. [87]	Sangeetha et al. [82]	Bardi et al. [88]	Alwan et al. [83]	Proposal
Collection	Privacy Policy					✓	✓											✓
	Privacy-Preserving Data Collection																	✓
Storages	Encryption	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Access Control	✓			✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Audit Trail				✓	✓							✓	✓	✓			✓
Analytics	Privacy Preserving Data Mining	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓
	Access Control					✓												✓
Utilization	Audit Trail			✓		✓												✓
	Privacy-Preserving Data Publishing		✓	✓								✓	✓	✓				✓
Destruction	Degaussing																	✓
	Overwriting																	✓

5. Discussion and Open Challenge

We previously analyzed the security-related standards trend in Section 2 and presented the taxonomy of security and privacy in Section 3. In addition, we conducted a comparative evaluation of the survey studies and our proposal in Section 4, and based on it, this section describes related issues and open challenges. In particular, we describe security issues that are difficult to fully address with technologies developed until recently and present open challenges as future work to be solved in the big data life cycle.

In the collection phase, data provenance is one of the issues of big data. If data is collected indiscriminately, the source of the data is unclear and noisy data is collected. This data affects the analytics phase and damages the reliability of the analysis. In addition, because a lot of unstructured data is collected, it is necessary to properly classify it. Therefore, in the collection phase, one of the major challenges is to properly filter and classify the data so as not to compromise its reliability. In addition, when the data provider provides the data, control of the data moves away from the provider. Therefore, it is necessary to ensure the rights of the data subject by ensuring that users know whether they are properly managed or used properly [79]. Since homomorphic encryption can be operated without decryption, a lot of research is being conducted in this sphere. In theory, homomorphic encryption can provide the best results for data miners while completely protecting the privacy of data providers. However, because of the high turnaround time required, it is only used in the collection and storage phases, and cannot be used in the analytics phase. Therefore, the high turnaround time of homomorphic encryption is one of the major challenges of the big data life cycle.

In the storage phase, the collected informal data must be stored properly. Because most big data are stored in a distributed environment, there is a possibility of lower query performance. In particular, because a lower performance affects the analytics phase, maintaining appropriate query performance and throughput is a major issue in the storage phase. In addition, many big data storage systems use existing cryptographic algorithms such as AES, RSA, and IBE. However, due to the characteristics of a distributed environment, key management issues arise and lead to lower performance. Therefore, it is required to a cryptographic algorithm suitable in a big data environment. Despite much research on storage auditing in the cloud, storage overhead, communication costs, and security challenges remain unresolved. In particular, the data leakage issue in collaborative auditing, and the replay attack weakness of specific storage auditing methods are security challenges that must be solved for a safe storage auditing method [49].

In the analytics phase, various privacy technologies are used to prevent exposure of the data subject. However, research on comparison tools (e.g., benchmarks) for evaluating and comparing the efficacy performance is lacking. There is a lot of research going on many PPDMS, but most of them are increasingly geared towards data utilities. Specific miners can identify individuals for data utility. Therefore, the balance of privacy and utility is a major challenge in PPDMS. In addition, security and privacy issues are considered in the limitations of specific mining methods. Therefore, it is not possible to generalize the data mining algorithm results, which is seen as a risk and open challenge to information disclosure [94]. Finally, users performing data processing and analysis should not be tied to a specific platform. Still, they should use a variety of processing platforms to achieve high efficiency and scalability. With relational databases such as PostgreSQL and traditional queries, they can aggregate large data sets. However, it can be faster to do in Spark for ML jobs [95].

The open challenge in the utilization phase is also related to balancing between utility and privacy. However, if the level of de-identification is lowered by prioritizing the utility of analytics, the data visualized in the utilization phase can expose confidential information in combination with other data. In addition, data mining can unintentionally expose sensitive data. To solve these problems, de-identification techniques for the utilization phase are needed. Visualization is especially used in the utilization phase, but there are not many studies related to privacy-preserving sensitive data. Therefore, the lack of suitable privacy-preserving technology in the utilization phase is a major challenge.

In the data destruction phase, there are issues related to distributed environments and web storage such as cloud systems. Currently, in acts of protecting privacy such as in GDPR, there are laws to destroy collected privacy. In addition, some standards also describe the big data life cycle and the requirements for data destruction. However, recent studies related to big data recognize the importance of data destruction, but do not focus on the development of related data destruction technologies. Existing overwriting and degaussing are used as techniques to destroy data, but they cannot be used in distributed environments and web storage such as the cloud. Most big data studies construct a cloud environment as the amount of collected data increases exponentially. Because the cloud is a distributed storage based on virtualization technology, the overwriting technique requires an accurate data path and a degaussing technique that disables physical storage. Therefore, novel data destruction techniques are required that are different from the existing big data destruction techniques in cloud environments.

6. Conclusions

Big data offers several advantages and promising potential for innovation in a variety of fields, but it also presents many issues and challenges. In particular, each phase of the big data life cycle has data security and reliability issues, and it can be a threat to privacy invasion through various big data analyses. Therefore, this paper identified threats and security issues arising in the life cycle of big data by confirming the current standards developed by international standards organizations and analyzing related studies. We also divided the big data life cycle into five phases (i.e., collection, storage, analytics, utilization, and destruction) and defined the security classification of the big data life cycle based on identified threats and security issues. Based on the comparative evaluation, we discussed the related issues and open challenges in the big data life cycle. Finally, we conducted an evaluation through comparison of our proposed security taxonomy against existing big data security and privacy-preserving research. Although security and privacy are important issues in big data, many standards organizations do not cover requirements and technologies in detail. Therefore, the current standard status and related studies we surveyed can be used to highlight the need for security and privacy in a big data environment. In addition, the security taxonomy we have classified can be used as a guideline for evaluating other big data life cycle-related studies. In most current studies, data collection and destruction studies of the big data life cycle have been less done than storage and analysis. However, because all phases in the big data life cycle are interrelated and have a great influence, security, and privacy issues at all phases should be addressed. In future work, we intend to clarify the proposed security taxonomy and design a security architecture according to the big data life cycle.

Author Contributions: The authors contributed to this paper as follows: J.K. wrote this article, classified and designed the taxonomy; G.K. analyzed the standards and coordinated the design of taxonomy and the analysis of papers; Y.-G.K. supervised and coordinated the investigation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant of the Korea Health Technology R & D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C1140).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. ISO—International Organization for Standardization. Available online: <https://www.iso.org/about-us.html> (accessed on 27 October 2020).
2. ITU Telecommunication Standardization Sector. Available online: <https://www.itu.int/en/ITU-T/about/Pages/default.aspx> (accessed on 27 October 2020).
3. ISO/IEC JTC1—Information Technology—ISO. Available online: <https://www.iso.org/isoiec-jtc-1.html> (accessed on 27 October 2020).

4. NIST: National Institute of Standards and Technology. Available online: <https://www.nist.gov/about-nist> (accessed on 27 October 2020).
5. SAC—Standardization Administration of China—ISO. Available online: <https://www.iso.org/member/1635.html> (accessed on 27 October 2020).
6. BSI—British Standards Institution—ISO. Available online: <https://www.iso.org/member/2064.html> (accessed on 27 October 2020).
7. TTA—Telecommunications Technology Association. Available online: <https://www.tta.or.kr/eng/index.jsp> (accessed on 27 October 2020).
8. TM Forum—How to manage Digital Transformation, Agile Business Operations & Connected Digital Ecosystems. Available online: <https://www.tmforum.org/> (accessed on 27 October 2020).
9. IEEE SA—The IEEE Standards Association—Home. Available online: <https://standards.ieee.org/> (accessed on 27 October 2020).
10. Apache Hadoop. Available online: <https://hadoop.apache.org/> (accessed on 27 October 2020).
11. Greene, T.; Shmueli, G.; Ray, S.; Fell, J. Adjusting to the GDPR: The impact on data scientists and behavioral researchers. *Big Data* **2019**, *7*, 140–162. [[CrossRef](#)] [[PubMed](#)]
12. Stallings, W. Handling of Personal Information and Deidentified, Aggregated, and Pseudonymized Information under the California Consumer Privacy Act. *IEEE Secur. Priv.* **2020**, *18*, 61–64. [[CrossRef](#)]
13. Kanika, A.; Khan, R.A. An Improved Security Threat Model for Big Data Life Cycle. *Asian J. Comput. Sci. Technol.* **2018**, *7*, 33–39.
14. Hornyack, P.; Han, S.; Jung, J.; Schechter, S.; Wetherall, D. These aren't the droids you're looking for: Retrofitting android to protect data from imperious applications. In Proceedings of the 18th ACM Conference on Computer and Communications Security, Chicago, IL, USA, 17–21 October 2011. [[CrossRef](#)]
15. Zhao, Y.; Wang, Z.; Zou, L.; Wang, J.; Hao, Y. A Linked Data Based Personal Service Data Collection and Semantics Unification Method. In Proceedings of the 2014 International Conference on Service Sciences, Wuxi, China, 22–23 May 2014. [[CrossRef](#)]
16. Gao, W.; Yu, W.; Liang, F.; Hatcher, W.G.; Lu, C. Privacy-preserving auction for big data trading using homomorphic encryption. *IEEE Trans. Netw. Sci. Eng.* **2020**, *7*, 776–791. [[CrossRef](#)]
17. Mittal, D.; Kaur, D.; Aggarwal, A. Secure data mining in cloud using homomorphic encryption. In Proceedings of the 2014 IEEE International Conference on Cloud Computing in Emerging Markets (CEEM), Bangalore, India, 15–17 October 2014. [[CrossRef](#)]
18. Balebako, R.; Jung, J.; Lu, W.; Cranor, L.F.; Nguyen, C. “Little brothers watching you”: Raising awareness of data leaks on smartphones. In Proceedings of the Ninth Symposium on Usable Privacy and Security, Newcastle, UK, 24–26 July 2013. [[CrossRef](#)]
19. Liu, S.; Qu, Q.; Chen, L.; Ni, L.M. SMC: A practical schema for privacy-preserved data sharing over distributed data streams. *IEEE Trans. Big Data* **2015**, *1*, 68–81. [[CrossRef](#)]
20. Gupta, A.; Verma, A.; Kalra, P.; Kumar, L. Big Data: A security compliance model. In Proceedings of the 2014 Conference on IT in Business, Industry and Government (CSIBIG), Indore, India, 8–9 March 2014. [[CrossRef](#)]
21. Al-Shomrani, A.; Fathy, F.; Jambi, K. Policy enforcement for big data security. In Proceedings of the 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), Abha, Saudi Arabia, 26–27 March 2017. [[CrossRef](#)]
22. Consolvo, S.; Jung, J.; Greenstein, B.; Powledge, P.; Maganis, G.; Avrahami, D. The Wi-Fi privacy ticker: Improving awareness & control of privacy exposure on Wi-Fi. In Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Copenhagen, Denmark, 26–29 September 2010. [[CrossRef](#)]
23. Zhou, Y.; Zhang, X.; Jiang, X.; Freeh, V.W. Taming information-stealing smartphone applications (on android). In Proceedings of the International Conference on Trust and Trustworthy Computing, Berlin/Heidelberg, Germany, 21–23 June 2011. [[CrossRef](#)]
24. Tiwari, P.K.; Velayutham, T. Detection and deterrence from data collecting applications in Android. In Proceedings of the 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Wagnaghat, India, 22–24 December 2016. [[CrossRef](#)]
25. Xu, S.; Yang, G.; Mu, Y.; Liu, X. A secure IoT cloud storage system with fine-grained access control and decryption key exposure resistance. *Future Gener. Comput. Syst.* **2019**, *97*, 284–294. [[CrossRef](#)]
26. Li, J.; Lin, X.; Zhang, Y.; Han, J. KSF-OABE: Outsourced attribute-based encryption with keyword search function for cloud storage. *IEEE Trans. Serv. Comput.* **2016**, *10*, 715–725. [[CrossRef](#)]

27. Xue, L.; Yu, Y.; Li, Y.; Au, M.H.; Du, X.; Yang, B. Efficient attribute-based encryption with attribute revocation for assured data deletion. *Inf. Sci.* **2019**, *479*, 640–650. [[CrossRef](#)]
28. Yang, P.; Xiong, N.; Ren, J. Data Security and Privacy Protection for Cloud Storage: A Survey. *IEEE Access* **2020**, *8*, 131723–131740. [[CrossRef](#)]
29. Baek, J.; Vu, Q.H.; Liu, J.K.; Huang, X.; Xiang, Y. A secure cloud computing based framework for big data information management of smart grid. *IEEE Trans. Cloud Comput.* **2014**, *3*, 233–244. [[CrossRef](#)]
30. Zhang, Y.; Yang, M.; Zheng, D.; Lang, P.; Wu, A.; Chen, C. Efficient and secure big data storage system with leakage resilience in cloud computing. *Soft Comput.* **2018**, *22*, 7763–7772. [[CrossRef](#)]
31. Azougaghe, A.; Kartit, Z.; Hedabou, M.; Belkasm, M.; El Marraki, M. An efficient algorithm for data security in cloud storage. In Proceedings of the 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), Marrakech, Morocco, 14–16 December 2015. [[CrossRef](#)]
32. Hussien, Z.A.; Jin, H.; Abduljabbar, Z.A.; Hussain, M.A.; Abbdal, S.H.; Zou, D. Scheme for ensuring data security on cloud data storage in a semi-trusted third party auditor. In Proceedings of the 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, China, 19–20 December 2015. [[CrossRef](#)]
33. Li, Y.; Gai, K.; Qiu, L.; Qiu, M.; Zhao, H. Intelligent cryptography approach for secure distributed big data storage in cloud computing. *Inf. Sci.* **2017**, *387*, 103–115. [[CrossRef](#)]
34. Al-Odat, Z.; Al-Qtiemat, E.; Khan, S. A big data storage scheme based on distributed storage locations and multiple authorizations. In Proceedings of the 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), Washington, DC, USA, 27–29 May 2019. [[CrossRef](#)]
35. Arora, A.; Khanna, A.; Rastogi, A.; Agarwal, A. Cloud security ecosystem for data security and privacy. In Proceedings of the 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, Noida, India, 12–13 January 2017. [[CrossRef](#)]
36. Saroj, S.K.; Chauhan, S.K.; Sharma, A.K.; Vats, S. Threshold cryptography based data security in cloud computing. In Proceedings of the 2015 IEEE International Conference on Computational Intelligence & Communication Technology, Ghaziabad, India, 13–14 February 2015. [[CrossRef](#)]
37. Bajwa, M.S.; Kang, S.S. An Enhanced Data Owner Centric Model for Ensuring Data Security in Cloud. In Proceedings of the 2015 second International Conference on Advances in Computing and Communication Engineering, Dehradun, India, 1–2 May 2015. [[CrossRef](#)]
38. Sanka, S.; Hota, C.; Rajarajan, M. Secure data access in cloud computing. In Proceedings of the 2010 IEEE 4th International Conference on Internet Multimedia Services Architecture and Application, Bangalore, India, 15–17 December 2010. [[CrossRef](#)]
39. Cheng, H.; Rong, C.; Hwang, K.; Wang, W.; Li, Y. Secure big data storage and sharing scheme for cloud tenants. *China Commun.* **2015**, *12*, 106–115. [[CrossRef](#)]
40. Al Hamid, H.A.; Rahman, S.M.M.; Hossain, M.S.; Almogren, A.; Alamri, A. A security model for preserving the privacy of medical big data in a healthcare cloud using a fog computing facility with pairing-based cryptography. *IEEE Access* **2017**, *5*, 22313–22328. [[CrossRef](#)]
41. Saraiva, D.A.; Leithardt, V.R.Q.; de Paula, D.; Sales Mendes, A.; González, G.V.; Crocker, P. Prisc: Comparison of symmetric key algorithms for IOT devices. *Sensors* **2019**, *19*, 4312. [[CrossRef](#)] [[PubMed](#)]
42. Ko, S.Y.; Jeon, K.; Morales, R. The *HybrEx* Model for Confidentiality and Privacy in Cloud Computing. *HotCloud* **2011**, *11*, 1–5. [[CrossRef](#)]
43. Ngo, C.; Membrey, P.; Demchenko, Y.; de Laat, C. Policy and context management in dynamically provisioned access control service for virtualized cloud infrastructures. In Proceedings of the 2012 Seventh International Conference on Availability, Reliability and Security, Prague, Czech Republic, 20–24 August 2012. [[CrossRef](#)]
44. Yu, S.; Wang, C.; Ren, K.; Lou, W. Achieving secure, scalable, and fine-grained data access control in cloud computing. In Proceedings of the 2010 Proceedings IEEE INFOCOM, San Diego, CA, USA, 15–19 March 2010. [[CrossRef](#)]
45. Younis, Y.A.; Kifayat, K.; Merabti, M. An access control model for cloud computing. *J. Inf. Secur. Appl.* **2014**, *19*, 45–60. [[CrossRef](#)]
46. Liu, Q.; Wang, G.; Wu, J. Secure and privacy preserving keyword searching for cloud storage services. *J. Netw. Comput. Appl.* **2012**, *35*, 927–933. [[CrossRef](#)]

47. Adrienne, F.; David, E. Privacy protection for social networking APIs. In Proceedings of the Web 2.0 Security and Privacy 2008 (In Conjunction with 2008 IEEE Symposium on Security and Privacy), Oakland, CA, USA, 22 May 2008.
48. Sundareswaran, S.; Squicciarini, A.; Lin, D. Ensuring distributed accountability for data sharing in the cloud. *IEEE Trans. Dependable Secur. Comput.* **2012**, *9*, 556–568. [[CrossRef](#)]
49. Yang, K.; Jia, X. Data storage auditing service in cloud computing: Challenges, methods and opportunities. *World Wide Web* **2012**, *15*, 409–428. [[CrossRef](#)]
50. Ferdous, M.S.; Margheri, A.; Paci, F.; Yang, M.; Sassone, V. Decentralised runtime monitoring for access control systems in cloud federations. In Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, USA, 5–8 June 2017. [[CrossRef](#)]
51. Wang, C.; Wang, Q.; Ren, K.; Lou, W. Ensuring data storage security in Cloud Computing. In Proceedings of the 2009 17th International Workshop on Quality of Service, Charleston, SC, USA, 13–15 July 2009. [[CrossRef](#)]
52. Mohan, S.V.; Angamuthu, T. Association Rule Hiding in Privacy Preserving Data Mining. *Int. J. Inf. Secur. Priv.* **2018**, *12*, 141–163. [[CrossRef](#)]
53. Gopalan, N.P.; Murthy, T.S. Association Rule Hiding Using Chemical Reaction Optimization. In Proceedings of the Soft Computing for Problem Solving, Bhubaneswar, India, 23–24 December 2018. [[CrossRef](#)]
54. Menaga, D.; Revathi, S. Least lion optimisation algorithm (LLOA) based secret key generation for privacy preserving association rule hiding. *IET Inf. Secur.* **2018**, *12*, 332–340. [[CrossRef](#)]
55. Liu, Y.; Guo, W.; Fan, C.L.; Chang, L.; Cheng, C. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. *IEEE Trans. Ind. Inform.* **2018**, *15*, 1767–1774. [[CrossRef](#)]
56. Gahar, R.M.; Arfaoui, O.; Hidri, M.S.; Hadj-Alouane, N.B. A Distributed Approach for High-Dimensionality Heterogeneous Data Reduction. *IEEE Access* **2019**, *7*, 151006–151022. [[CrossRef](#)]
57. Motiwalla, L.; Li, X. Value added privacy services for healthcare data. In Proceedings of the 2010 6th World Congress on Services, Miami, FL, USA, 5–10 July 2010. [[CrossRef](#)]
58. Cui, B.; Zhang, B.; Wang, K. A data masking scheme for sensitive big data based on format-preserving encryption. In Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, China, 21–24 July 2017. [[CrossRef](#)]
59. Gao, Z.Q.; Zhang, L.J. DPHKMS: An efficient hybrid clustering preserving differential privacy in spark. In Proceedings of the International Conference on Emerging Internetworking, Data & Web Technologies, Wuhan, China, 10–11 June 2017. [[CrossRef](#)]
60. Ni, L.; Li, C.; Wang, X.; Jiang, H.; Yu, J. DP-MCDBSCAN: Differential privacy preserving multi-core DBSCAN clustering for network user data. *IEEE Access* **2018**, *6*, 21053–21063. [[CrossRef](#)]
61. Mo, R.; Liu, J.; Yu, W.; Jiang, F.; Gu, X.; Zhao, X.; Liu, W.; Peng, J. A Differential Privacy-Based Protecting Data Preprocessing Method for Big Data Mining. In Proceedings of the 2019 18th IEEE International Conference On Trust, Security And Privacy in Computing and Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 5–8 August 2019. [[CrossRef](#)]
62. Zhao, L.; Ni, L.; Hu, S.; Chen, Y.; Zhou, P.; Xiao, F.; Wu, L. Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications, Honolulu, HI, USA, 15–19 April 2018. [[CrossRef](#)]
63. Lin, C.; Song, Z.; Song, H.; Zhou, Y.; Wang, Y.; Wu, G. Differential privacy preserving in big data analytics for connected health. *J. Med. Syst.* **2016**, *40*, 97. [[CrossRef](#)]
64. Zhang, X.; Yang, L.T.; Liu, C.; Chen, J. A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. *IEEE Trans. Parallel Distrib. Syst.* **2013**, *25*, 363–373. [[CrossRef](#)]
65. Mehta, B.B.; Rao, U.P. Privacy preserving big data publishing: A scalable k-anonymization approach using MapReduce. *IET Softw.* **2017**, *11*, 271–276. [[CrossRef](#)]
66. Machanavajhala, A.; Kifer, D.; Gehrke, J.; Venkatasubramanian, M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 1–12. [[CrossRef](#)]
67. Li, N.; Li, T.; Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 11–15 April 2007. [[CrossRef](#)]

68. Vatsalan, D.; Christen, P.; Verykios, V.S. A taxonomy of privacy-preserving record linkage techniques. *Inf. Syst.* **2013**, *38*, 946–969. [[CrossRef](#)]
69. Scannapieco, M.; Figotin, I.; Bertino, E.; Elmagarmid, A.K. Privacy preserving schema and data matching. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 11–14 June 2007. [[CrossRef](#)]
70. Gkoulalas-Divanis, A.; Verykios, V.S. Exact knowledge hiding through database extension. *IEEE Trans. Knowl. Data Eng.* **2008**, *21*, 699–713. [[CrossRef](#)]
71. Verykios, V.S.; Elmagarmid, A.K.; Bertino, E.; Saygin, Y.; Dasseni, E. Association rule hiding. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 434–447. [[CrossRef](#)]
72. Verykios, V.S. Association rule hiding methods. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 28–36. [[CrossRef](#)]
73. Dasgupta, A.; Kosara, R. Adaptive privacy-preserving visualization using parallel coordinates. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2241–2248. [[CrossRef](#)] [[PubMed](#)]
74. Dasgupta, A.; Maguire, E.; Abdul-Rahman, A.; Chen, M. Opportunities and challenges for privacy-preserving visualization of electronic health record data. In Proceedings of the IEEE VIS 2014 Workshop on Visualization of Electronic Health Records, Paris, France, 9–14 November 2014. [[CrossRef](#)]
75. Chou, J.K.; Wang, Y.; Ma, K.L. Privacy preserving event sequence data visualization using a Sankey diagram-like representation. In Proceedings of the SIGGRAPH ASIA 2016 Symposium on Visualization, Macao, China, 5–8 December 2016. [[CrossRef](#)]
76. Toysmart.com, LLC, and Toysmart.com, Inc. | Federal Trade Commission. Available online: <https://www.ftc.gov/enforcement/cases-proceedings/x000075/toysmartcom-llc-toysmartcom-inc> (accessed on 27 October 2020).
77. PIPEDA Case Summary 2003-189: Bank Removed Customer’s SIN from some, but not all, of Its Records—Office of the Privacy Commissioner of Canada. Available online: <https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2003/pipeda-2003-189/> (accessed on 27 October 2020).
78. Abouelmehdi, K.; Beni-Hessane, A.; Khaloufi, H. Big healthcare data: Preserving security and privacy. *J. Big Data* **2018**, *5*, 1. [[CrossRef](#)]
79. Yu, S. Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE Access* **2016**, *4*, 2751–2763. [[CrossRef](#)]
80. Ye, H.; Cheng, X.; Yuan, M.; Xu, L.; Gao, J.; Cheng, C. A survey of security and privacy in big data. In Proceedings of the 2016 16th International Symposium on Communications and Information Technologies (ISCIT), Qingdao, China, 26–28 September 2016. [[CrossRef](#)]
81. Wang, T.; Zheng, Z.; Rehmani, M.H.; Yao, S.; Huo, Z. Privacy preservation in big data from the communication perspective—A survey. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 753–778. [[CrossRef](#)]
82. Sangeetha, S.; Sadasivam, G.S. Privacy of big data: A review. In *Handbook of Big Data and IoT Security*; Springer: Cham, Switzerland, 2019; pp. 5–23. [[CrossRef](#)]
83. Alwan, H.B.; Ku-Mahamud, K.R. Big data: Definition, characteristics, life cycle, applications, and challenges. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Pahang, Malaysia, 25–27 September 2019. [[CrossRef](#)]
84. Xu, L.; Jiang, C.; Wang, J.; Yuan, J.; Ren, Y. Information security in big data: Privacy and data mining. *IEEE Access* **2014**, *2*, 1149–1176. [[CrossRef](#)]
85. Jain, P.; Gyanchandani, M.; Khare, N. Big data privacy: A technological perspective and review. *J. Big Data* **2016**, *3*, 25. [[CrossRef](#)]
86. Lv, D.; Zhu, S.; Xu, H.; Liu, R. A Review of Big Data Security and Privacy Protection Technology. In Proceedings of the 2018 18th International Conference on Communication Technology (ICCT), Chongqing, China, 8–11 October 2018. [[CrossRef](#)]
87. Goswami, P.; Madan, S. A survey on big data & privacy preserving publishing techniques. *Adv. Comput. Sci. Technol.* **2017**, *10*, 395–408.
88. Bertino, E.; Ferrari, E. Big data security and privacy. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*; Springer: Cham, Switzerland, 2018; pp. 425–439. [[CrossRef](#)]
89. Fang, W.; Wen, X.Z.; Zheng, Y.; Zhou, M. A survey of big data security and privacy preserving. *IETE Tech. Rev.* **2017**, *34*, 544–560. [[CrossRef](#)]

90. Jiang, H.; Wang, K.; Wang, Y.; Gao, M.; Zhang, Y. Energy big data: A survey. *IEEE Access* **2016**, *4*, 3844–3861. [[CrossRef](#)]
91. Mehmood, A.; Natgunanathan, I.; Xiang, Y.; Hua, G.; Guo, S. Protection of big data privacy. *IEEE Access* **2016**, *4*, 1821–1834. [[CrossRef](#)]
92. Alshboul, Y.; Nepali, R.; Wang, Y. Big data lifecycle: Threats and security model. In Proceedings of the Twenty-first Americas Conference on Information Systems (AMCIS), Fajardo, Puerto Rico, 13–15 August 2015; ISBN 978-0-9966831-0-4.
93. Moreno, J.; Serrano, M.A.; Fernández-Medina, E. Main issues in big data security. *Future Internet* **2016**, *8*, 44. [[CrossRef](#)]
94. Verykios, V.S.; Bertino, E.; Fovino, I.N.; Provenza, L.P.; Saygin, Y.; Theodoridis, Y. State-of-the-art in privacy preserving data mining. *ACM SIGMOD Rec.* **2004**, *33*, 50–57. [[CrossRef](#)]
95. Agrawal, D.; Chawla, S.; Elmagarmid, A.K.; Kaoudi, Z.; Ouzzani, M.; Papotti, P.; Quiané-Ruiz, J.A.; Tang, N.; Zaki, M.J. Road to Freedom in Big Data Analytics. In Proceeding of the 19th International Conference on Extending Database Technology (EDBT), Bordeaux, France, 15–18 March 2016. [[CrossRef](#)]

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).