Original Research

SAGE Open
April-June 2022: 1–23
© The Author(s) 2022
DOI: 10.1177/21582440221096445
journals.sagepub.com/home/sgo
SAGE

Privacy Prevention of Big Data Applications: A Systematic Literature Review

Fatima Rafiq[1], Mazhar Javed Awan[1], Awais Yasin[2], Haitham Nobanee[3], Azlan Mohd Zain[4], and Saeed Ali Bahaj[5]

Abstract

This paper focuses on privacy and security concerns in Big Data. This paper also covers the encryption techniques by taking existing methods such as differential privacy, *k*-anonymity, *T*-closeness, and *L*-diversity. Several privacy-preserving techniques have been created to safeguard privacy at various phases of a large data life cycle. The purpose of this work is to offer a comprehensive analysis of the privacy preservation techniques in Big Data, as well as to explain the problems for existing systems. The advanced repository search option was utilized for the search of the following keywords in the search: "Cyber security" OR "Cybercrime") AND (("privacy prevention") OR ("Big Data applications")). During Internet research, many search engines and digital libraries were utilized to obtain information. The obtained findings were carefully gathered out of which 103 papers from 2,099 were found to gain the best information sources to address the provided study subjects. Hence a systemic review of 32 papers from 103 found in major databases (IEEExplore, SAGE, Science Direct, Springer, and MDPIs) were carried out, showing that the majority of them focus on the privacy prediction of Big Data applications with a contents-based approach and the hybrid, which address the major security challenge and violation of Big Data. We end with a few recommendations for improving the efficiency of Big Data projects and provide secure possible techniques and proposed solutions and model that minimizes privacy violations, showing four different types of data protection violations and the involvement of different entities in reducing their impacts.

Keywords

privacy, cyber security, anonymity, data protection, sustainability, Big Data, security, prevention, public policy, cybercrime, artificial intelligence, Internet of Things

Introduction

The phrase "Big Data" refers to the vast and ever-increasing volumes of data that might overwhelm an organization (Ur Rehman et al., 2016). It gathers massive, broad, and multi-format data streams from disparate and independent data sources (X. Wu et al., 2014). Big Data is believed to have five properties, which are known as the five V's: volume, velocity, variety, veracity, and valence (Ahmed et al., 2021). As illustrated in Figure 1, an overlapping sixth V has been added to the large data formula: value (Awan, Rahim, Nobanee, Yasin, et al., 2021).

When the capacity is obtained in large data centers and storage area networks, Big Data is the key property of massive data, reflected in their volume. The huge number of large data leads to data heterogeneity and a broad variety of dimensionalities in datasets. Efforts are therefore needed to decrease the amount to analyze large figures effectively (Che et al., 2013). In order to prevent the consumption of lateral storage and processing resources, massive data streams must be treated online. The speed of Big Data is the second key feature. The speed refers to the frequency of data streams must be lowered to handle huge data correctly. The observatory for solar dynamics creates about a terabyte of data each day, for example, and analysis of such rapid data can only be conceived after reduction and summary (Battams, 2015). The "curse of dimensionality" afflicts Big Data. To put it another way, in order to uncover the most knowledge patterns, millions of dimensions

[1]University of Management and Technology, Lahore, Pakistan
[2]National University of Technology, Islamabad, Pakistan
[3]Abu Dhabi University College of Business Administration, UAE
[4]Universiti Teknologi Malaysia, Skudai, Malaysia
[5]Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia

Corresponding Author:
Haitham Nobanee, Abu Dhabi University College of Business Administration, Al Ain Road, Abu Dhabi 59911, UAE.
Email: nobanee@gmail.com

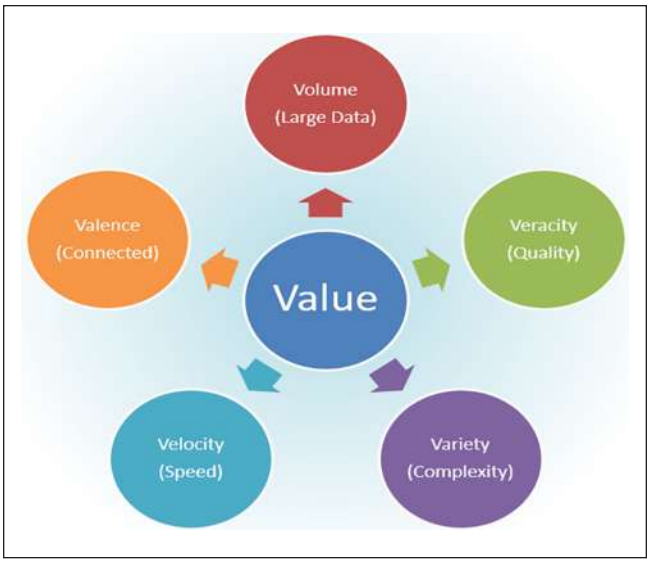

Figure 1. The six V's of Big Data.

(variables, traits, and attributes) must be successfully reduced (Zhai et al., 2014). Internet user activity profiles, for example, are sparse and vast, with millions of keywords and URLs that might be utilized (Chandramouli et al., 2012). Similarly, a personal high-performance genetic sequence not only increases data volume and speed, but it also contributes to the data's high dimensionality (Ward et al., 2013). The volume and diversity of large data are rising enormously for a variety of reasons. Big Data is also employed in several nations to provide services for fields such as health care, medicine, operations in the public sector, distribution, marketing, and production (Koo et al., 2020). Retailers, for example, are accumulating enormous databases about their customers' sales activity. Organizations are working on logistics and financial services, and social media users are sharing a lot of information on sales prices and items (Sánchez-Moreno et al., 2020). Volume and diversity in both organized and unstructured data are two of Big Data's problems (Awan, Rahim, Nobanee, Munawar, et al., 2021). The use of Big Data in higher education research (HE) improves education and learning processes, advances individual assessments and education systems, and optimizes decision-making and education governance (Florea & Florea, 2020).

These new methods to analytics might be quite transformative. Retail purchasers utilize Big Data Analytics, for example, frequently to identify products usually available during each season and to predict geographical locations where strong demand is expected. In addition to all the amazing business opportunities, Big Data analytics has so many advantages that it has also generated several new data protection issues (Chen & Lin, 2014).

Big Data analysis may be utilized on the basis of safety analytics for the identification of anomalies and fraud. The analysis could be carried out in a dispersed setting, as data is not always transferred from organization to organization. As so many kinds of data are maintained in many different systems, additional research such as information mining and statistical analysis can take place with the appropriate infrastructure for the study of huge data (Gai et al., 2016). On the one hand, massive amounts of data are collected and stored. while, on the other, huge amounts of data are more difficult to secure against unauthorized access (D. Zhang, 2018).

The cyber world also plays a significant role in rendering cyber threats and attacks indefinitely susceptible. Cybersecurity refers to the tactics, tools, and processes used to combat cyber-attacks and cyber-based threats (Craigen et al., 2014). In identifying and fighting fraud, cyber assaults and other dangers, traditional security solutions weren't particularly good. Intelligent cyber hackers may easily violate the operations of a company in order to gain sensitive data, such as intellectual property, number of credit card and customer databases, to hurt the firm (Varshney et al., 2020). The Internet of Things (IoT)-based applications are quickly expanding in today's society (Krishna et al., 2021). User data privacy become a prominent subject for IoT academic researchers (Guo et al., 2021). Electronic Health System Studies in IoT Cloud have developed effective security procedures that offer complete frames with software features of vital importance to ensure that data transmissions are secure and correct between devices (Butpheng et al., 2020). As the Internet becomes more affordable, the number of Internet users grows on a daily basis. As a result, the volume of data exchanged via the Internet is expanding (Awan, Khan, et al., 2021). Social networking is rapidly expanding on a daily basis these days. Every day, millions of people register for social media networks. People utilize social media for a variety of reasons, such as entertainment, education, and business (Awan et al., forthcoming). One of the most significant breakthroughs in recent years has been the greater mobility of portable yet powerful wireless devices capable of connecting across a wide range of wireless networks with varying link-level characteristics (Adjei et al., 2020). In the world of technology, retrieving information from one location to another is a critical component of the software world, which is the foundation of all technology, as there is no such technology that does not require information (Science, 2018). Almost every SNS is designed for information shares, uploads, views, downloads, and understandings. Common reasons for sharing information are to draw attention, build social capital, reinforce links between people, attract the same mind, and promote knowledge-based societies and information (Sharif et al., 2021). As previously stated, different people utilize it for different objectives, while some hackers or unauthorized individuals use these platforms for improper purposes by creating phony profiles. They construct a bogus profile with a bogus photo by impersonating another person. They utilize these profiles to disseminate false information, defraud others, and engage in a variety of other illegal activities (Awan, 2020). Cybercrime is rising

quicker than the current cybersecurity system in today's computer and IT environment. Some of the causes of vulnerabilities in a computer system to threat (Patel et al., 2008) include poor system configuration, a lack of expertise, and a restricted number of methods. Due to the rising cyber dangers, there must be greater progress while creating techniques for cybersecurity. Cyber dangers has significantly increased. The pace of safety hazards and the necessity to address these is becoming increasingly demanding (Peter, 2005). One of the most advanced methods for identifying cybercrime is machine learning. In order to meet the restrictions of traditional detection approaches, machine learning methods may be used (Firdausi et al., 2010). ML applications in many life sectors, including as education, health, business, and cybersecurity (Yasir et al., 2020) are being expanded (Jusas & Samuvel, 2019; Manjula & Anandaraju, 2018; Shaukat Dar & Ulya Azmeen, 2015).

The major aim of this article is to discuss data protection and security concerns and the ways to discover any shortcomings. This article discusses a number of Big Data Analytics security and privacy issues. In addition to commonly used security data sets, this research provides a brief overview of machine learning methods. Despite the fact that cyber security is paramount, there are breaches and roadblocks. This work also highlights the significant problems and constraints that face the use of cybersecurity (Shaukat et al., 2020).

The article is organized as follows: Section 2 is a review of the literature. Section 3 explains how the method used in the systematic review was developed; Section 4 provides discussion; and lastly Section 5 covers findings and about future work.

Literature Review

Today many of big data issues in privacy arise from a lack of knowledge on how to evaluate and prevent security issues of big data applications. There are various techniques to anonymization and algorithms available, as well as the possible solutions to protect user's privacy and violations that arise in Big Data systems. This section describes how Big Data can assist with security of information. Big Data analysis tools are most commonly used for analyzing and storing trend numbers for commercial reasons. Big Data analytical methods and technology are used in different industries such as insurance, health, credit cards, net banking, etc. (Mujahid et al., 2021). Big Data analytics can help detect fraud and doubtful behaviors through the examination of network traffic, financial transactions, and log files, along with the combination of many data sources into a logical view. But forensics and invasion detection have typically been a significant issue in studying log files, network routes, and system activities. Advanced large-scale data technology, such as Hadoop related databases and stream processing has been proven to store and analyze huge volumes of heterogeneous data on an unexpected scale.

Due to the huge potential of Big Data Analysis, in addition to all economic advantages, there has been a torrent of new data protection problems. Some of the worst privacy problems are as follows:

1. Anonymization may be impossible to achieve.
2. The analytics of bid data aren't completely accurate.
3. The individuals engaged are legally protected.
4. Unethical acts based on security intelligence interpretation and audit compliance
5. Incidents involving data breaches and fraud
6. Discrimination
7. Information security is a major data issue
8. Data masking might be denied if personal information is to be disclosed
9. Big Data will almost certainly die out
10. E-discovery concerns
11. Patents and copyrights will become obsolete

Information security is a Big Data issue. The research community is focusing on the development in the period of Big Data, computer science, and increasing business applications of quick and efficient algorithms for Big Data security intelligence, with the primary aim of ensuring a safe environment free of unlawful access (Cheng et al., 2017).

Security analysis will revolutionize these technologies by collecting and conducting comprehensive analysis of large scale data from a number of external and internal firm sources, such as susceptibility databases. Provide a full view of security data and analyze data in real time. Note that analysts and system architects still need to be informed on their systems to characterize "Big Data Tools" (Dev Mishra & Beer Singh, 2017).

Big Data analysis offers a great deal of promise for the improvement of all business areas, important advances and support for people in a number of ways for privacy issues. Organizations that adopt Big Data Analytics should, on the other hand, first consider the privacy and security issues before using analytics (Barth-Jones, 2012).

The key ideas and procedures for large data safety and confidentiality were explored during Bertinoet and Ferrari's in addition, they underlined major research concerns that need to be dealt with to ensure complete data safety and confidentiality solutions in the context of Big Data (Varshney et al., 2020).

Q. Zhang et al. (2016) established a computing paradigm that protects your privacy, which operated through the download of costly cloud activities. They have conceived a highly scalable technique for encryption and decryption, on the client side, while important processing activities are allocated to the cloud. Their technique increases preparation efficiencies 2.5 times while keeping private data confidential compared to traditional deep calculation.

Brki'c suggested a machine vision pipeline to assist preserve the privacy of people in video streams by keeping the

natural nature and utility of unidentifiable data and blurred faces (Kim & Park, 2020).

In his work Zhang (Yu, 2016) points out that Big Data gives both comfort and the privacy of individuals. He gives a quick overview of the problems posed by data collection, storage, and analysis. It highlights the necessity of validating the trustworthiness of Big Data as well as the weaknesses of existing technologies for data security and the legal side of data protection.

In their paper, Bertino and Ferrari express concerns on the threat posed to Big Data gathering by new data collecting and processing. In this paper, they investigate key ideas and approaches to protect the safety and confidentiality of Big Data against dangers such as these. They also underlined the problems of research to be solved to give the ever-growing amount of Big Data with full data security and privacy solutions (Varshney et al., 2020).

Xiang et al. (2016) formalized the general architecture of Big Data Analytics to address the new security and privacy challenges of data mining, recognized the conforming confidentiality requirements, and presented a cost-effective and privacy-preserving protocol for computational co-ordination as an example.

(Kantarcioglu & Shaon, 2019).

Kantarcioglu and Shaon (2019) has created a data protection and confidentiality solution for a uniform data defense in several data administration systems by complying with the safety criteria with the usage of advanced SECUREDL software. Companies can use their system to monitor sensitive data access, observe audit logs, sterilize, and accumulate sensitive data on the basis of data sensitivity and the needs of AI Models, notice illegal access, and develop data sensitivity and data type-based attribute-based access control policies.

Abouelmehdi et al. (2018) stressed that Big Data in the health care sector is important and beneficial. In their paper they conducted an investigation into current security and privacy problems in the healthcare field and evaluated exactly how safety and confidentiality issues in Big Data influence healthcare. Various data mining techniques were applied for medical treatment such as; decision tree *K*-Nearest Neighbor (KNN) Naive Bayes (Ali et al., 2019; Onan, 2015). Machine learning and deep learning techniques are being applied in various domains for-instance in health sector to improve the performance of technology (Awan, Raza, et al., 2021).

Deep learning techniques are being employed on large scale data for instance natural language processing domain is being used to analyze the general behavior through sentiment classification (Onan, 2017, 2021; Onan & Korukoğlu, 2016, 2017).

Many diseases can be predicted through machine learning and deep learning models (Aftab et al., 2021; Javed et al., 2021). There are additional regression models used to evaluate healthcare, notably polynomial regression, decision trees regression, and random forest regression (Awan et al., 2019; Awan, Rahim, Salim, et al., 2021; Gupta et al., 2021). For security point of view, they should be reviewed and evaluated the anonymization and coding techniques, their advantages and disadvantages lately offered and proposed future guidance for study.

D. Wu et al. (2016) described the application of the large-scale Wireless Sensor Networks (WSNs) method to Scalable Privacy Preservation (Sca-PBDA; WSNs). They endorsed their suggested technology with simulation results showing that it decreased the use of network resources and maintaining the privacy of sensors.

In this study, Gurajala et al. (2015) addressed how online social media platforms such as Twitter and Facebook have become highly popular for communication. Many businesses and people use it for various purposes. Twitter has grown highly popular with young adults and government users as a method of instantly interacting with their audience and simply communicating their ideas. . Tweets as well as some other communication data is being used in research for sentiment analysis and sarcasm detection (Onan, 2019; Onan & Tocoglu, 2020, 2021).

The cloud data via cloud operators has been emphasized by Gai et al. (2016). This jeopardizes users' cloud data. They presented a new technique in their study for effectively dividing a file and separately storing files on the cloud. Their preferred choice is Security-Aware Distributed Storage (SAEDS), which is primarily supported by Secure Efficient Data Distribution (SED2) and Efficient Data Conflation (EDCon) algorithms. They also assessed the safety and efficiency of the system.

Ambalavanan (2020) proposed various techniques for efficient detection of cyber pressures. One of the main disadvantages of the safety system is that computer resources typically decide the safety dependability levels of ordinary users with no technical safety expertise. A spam message also poses a threat to workstation resources. Junk messages are undesired and requested communications, which use a lot, together with computer memory and speed, of network resources. ML technology is used for the detection and classification of a communication as spam or ham. There are other various ML techniques that can forecast data (Applications et al., n.d.). ML-techniques are significantly involved in detecting computer spam communications (Chandrasekar, 2018), mobile text messages (Abdulhamid et al., 2017), spam tweets, and video pictures. The IDS is a computer network security system for scanning network vulnerabilities against malicious invasions. Signature, anomaly, and hybrid based detection system are the main classes of an interruption detection system for network analyses.

Machine Learning technology has a significant impact on the detection of different forms of network and host computer breaches. Numerous sectors are, nevertheless, considered as important problems for ML methods as zero-day detection and novel assaults (Jusas et al., 2019). Cyber defense systems may use a combination of approaches to prevent data breaches. For servers that store and process

Table 1. Research Queries.

Sr. No	Research interrogations	Main enthusiasm
RQ1	What are the privacy and security issues of Big Data?	To identify what types of issues, Big Data faced in terms of privacy and security
RQ2	What are the possible solutions/ measures may be put in place to protect user's privacy and security?	To investigate the possible solutions and procedures that can be taken to protect user's privacy.
RQ3	What are the different forms of privacy violations that arise in Big Data systems?	To determine the impact of privacy infractions in large data systems.
RQ4	What are encryption techniques for various activities in Big Data analytics?	To identify the anonymization techniques for Big Data systems
RQ5	What are the suggestions for restricting privacy violations?	To discover the types of risks of information security in Big Data applications

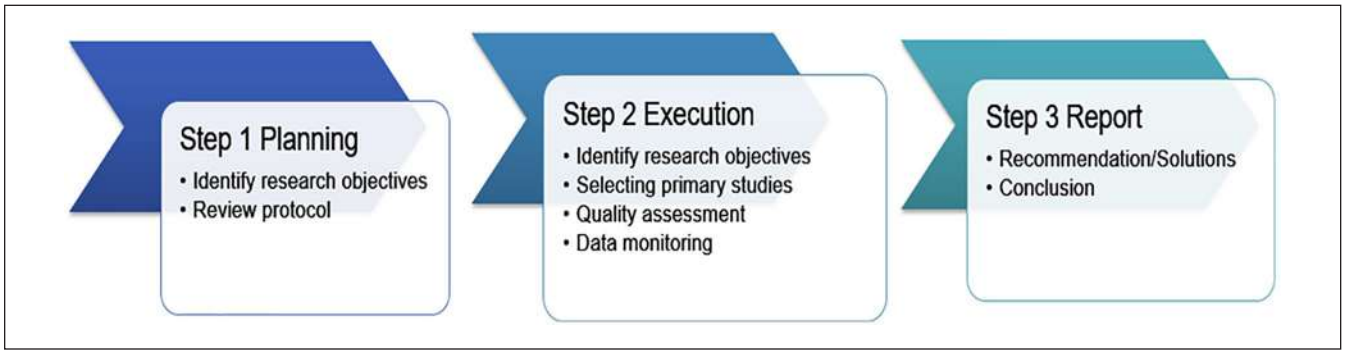

Figure 2. Stages of systematic review of literature.

data, physical security is essential. Honeypots and other spy systems are used for enhancing data security along with preventive techniques such as firewalls. In order to detect malicious activities access logs and alarm systems. In data storage and transmission, encoding techniques are also used. Despite these safeguards, data infringements continue to be widespread and devastating and cyber criminals have identified new ways of fighting (Shamsi & Khojaye, 2018).

Research Methodology

This Systematic Literature Review is designed to identify any privacy gap in Big Data applications, in particular in cybercrimes and various areas of the Internet, aspects related to the privacy of Big Data applications, such as techniques used to identify the privacy gaps and the security issues of the Big Data review, which may lead to solutions for protecting users' privacy. As a result, the following search queries were established, with Table 1 detailing the purpose of each of them. We carry out a Systematic Literature Review, as illustrated in the diagram below in Figure 2.

Review Protocol

Once the objectives were defined and search queries were developed, the following repositories were established:

Search Process

The two primary digital libraries in the ground of Big Data Analytics were the IEEE, SAGE, and Science Direct, and the search was supplemented by Springer, MDPI, which are two databases with access to a broad range of applications in a variety of areas. The keywords were then chosen on the basis of the preliminary mapping and the main terms found in their abstracts.

The terms utilized were: Data protection prevention, Big Data analysis, cybercrime, safety, and cyber security. The advanced repository option was utilized for the search by the use of the following in the search: "Cyber security" OR "Cybercrime") AND (("privacy prevention") OR ("Big Data applications")), in order to adjust the results.

Search Process

This research comprised on the following objectives:

O1: To discuss the privacy and security concerns of Big Data
O2: The general goal of our study is to learn more about solutions for large data analysis problems
O3: Discuss different forms of privacy violations that arise in Big Data systems
O4: Identify different encryption techniques for various activities in Big Data analytics

Table 2. Search String.

Sources	Search string	Perspective
IEEE, SAGE, Springer, MDPI, and Science direct	(“Cyber security” OR “Cybercrime”) AND ((“privacy prevention”) OR (“Big Data applications”))	Big Data Application

Table 3. Keywords Used in Research Paper.

Sr. No	Index terms
1	Big Data, cyber-crime, and data analytics
2	Big Data applications, privacy prevention, and IOT
3	Cyber-security, security, and artificial intelligence

Table 4. Details of the Search.

Source	Search chain	Records
IEEE	((“big data privacy” “AND security”)) AND (“privacy” and “prevention”)	1,975
SAGE	((“Big data”)) AND (“privacy and prevention”)	10
Science Direct	TITLE - (“Security and Privacy”) AND (“big data”)	70
Springer	“Big data”	23
MDPI	(“Privacy and challenges of big data”) AND (“cybercrime”)	21
Total records		2,099

Research Questions

The general goal of our study is to learn more about solutions for large data analysis problems. The comprehensive mapping study covers various research issues to obtain an insight into this subject (RQ). The following five RQs with their motivations are presented in Table 1. These questions help us categorize current research into Big Data Analysis and suggest future field research possibilities.

Table 5. Research Resource.

Sr. No	Research resources
1	IEEE Explore
2	Science Direct
3	Springer
4	MDPI
5	SAGE

Search String

Another phase of Systematic Literature Review consists of searching for suitable research studies. For collecting published articles on research topics, a search string was set up. We did a pilot search using the precise terms, and we chose to limit the privacy prevention search string to only Big Data applications. In the pilot search, however, we also leveraged cyber security via Big Data and machine learning. During Internet research, many search engines and digital libraries were utilized to obtain information. The obtained findings were carefully gathered in order to gain the best information sources to address the provided study subjects. Selected search engines and digital libraries were chosen based on their scientific content and relevance to the objective of this work. The databases used were Science Direct, SAGE, IEEE xplore, Springer, and MDPI. The next step is to define uniform methods and search terms for search engines and digital libraries for technical and scientific literature. Table 2 shows the terms chosen from the study questions to define the search string.

Search Keywords

During the early stages of our investigation, we used a methodology to identify keywords. We initially collected key terms from our research questions in order to relate our research questions to our goals. The initial searches were to assist us adjust the selection of our keywords. In the section Keywords is a list of keywords indicated in Table 3.

Study Selection Process

The technique for searching and selecting study material was carried out in four stages.

1. A search for papers in electronic databases including the keywords “Big Data,” “Big Data analytics,” or “Big Data applications,” as well as “privacy prevention,” “cyber security,” or “cyber-crime.”
2. Examination of the title, abstract, and keywords of recognized publications, as well as selection of significant articles based on selection criteria.
3. Review of publications that were not deleted during the previous phase.
4. Scanning cross-reference articles for in-depth research.

Table 4 shows the detail of search string of each databases. When executing the search of the sources chosen, a total of 2,099 records were collected, the distribution of which may be shown in Table 5.

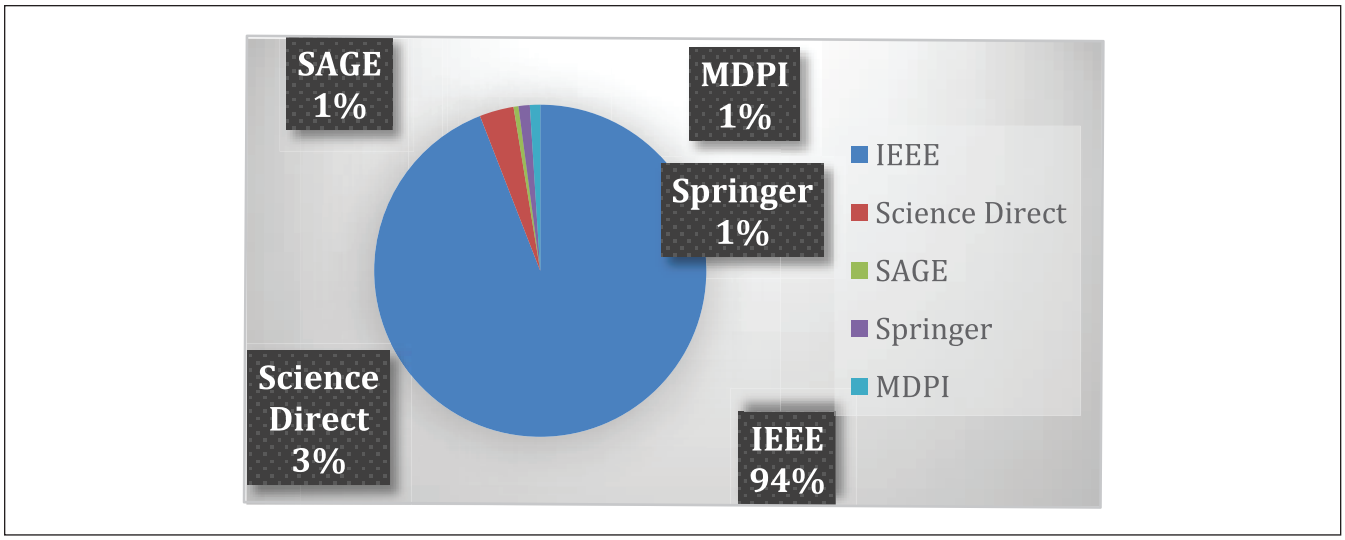

Figure 3. Graph depicting the details of the search string.

Figure 3 shows the detail of search string of each databases.

Research Resources and Methods

Data bases were utilized to locate the article in IEEE digital library, Springer, MDPI, and Science Directory. Also utilized by Google Scholars was to identify grey literature, such as white papers and technical studies. Google Scholar has shown itself to be a valuable tool for studies on bibliometric. The autocomplete function was triggered when the following search sentence was entered. Table 5 provides the search data basis for our study project literature.

Methods of research or methods, processes, and techniques for gathering data or evidence for analysis in order to disclose new data or acquire a better understanding of a topic are depicted in Table 6. This review article used the following research techniques, which are listed in Table 6 given below.

Figures 4 and 5 highlights the systematic review, and displays the number of articles by type of publication.

Result of Primary Studies

Once the research technique and topic have been determined, primary research is carried out utilizing keywords. The following primary research findings are given in Table 7 and graphically explained in Figure 6.

Figure 6 indicates the number of records of search string of each data-bases.

From the information in Table 4 and Figure 6, the areas with the biggest publication numbers are Asia, America, and China, as illustrated in Figure 7, based on the geographical distribution of the systemic review articles.

Table 6. Research Methods.

Sr.No	Research methods	Type of data used
1	Journals	Primary
2	Conferences	Primary

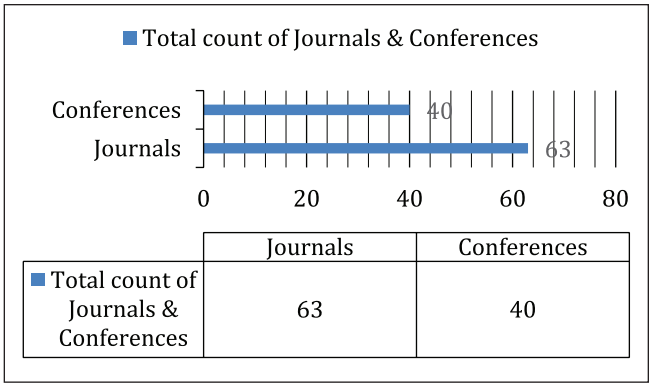

Figure 4. Count of journals and conferences in the systematic review.

Figure 8 displays the flow charts that indicate the number of records included and excluded in each selection step and the criteria for which major research should be included in the review.

Inclusion/Exclusion Criteria

After identifying these initial records, a first filter was conducted, excluding any duplicate experiments, yielding 1,537, to which the following inclusion and exclusion criteria were applied until 502 were obtained, as indicated in Table 8.

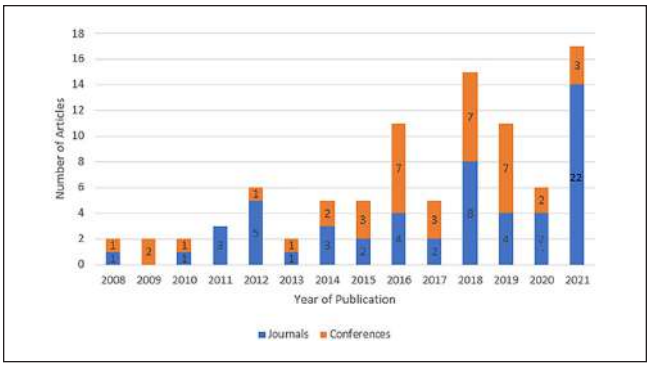

Figure 5. Distribution of the selected number of papers in the systematic review with respect to time of publication.

Table 7. Result of Primary Studies.

Sr.No	Search resource	Search date	Initial research results
1	IEEE Explore	20-June-2021	1,975
2	MDPI	27-July-2021	21
3	Science Direct	31-Aug-2021	70
4	Springer	1-Sep-2021	23
5	SAGE	31-Oct-2021	10

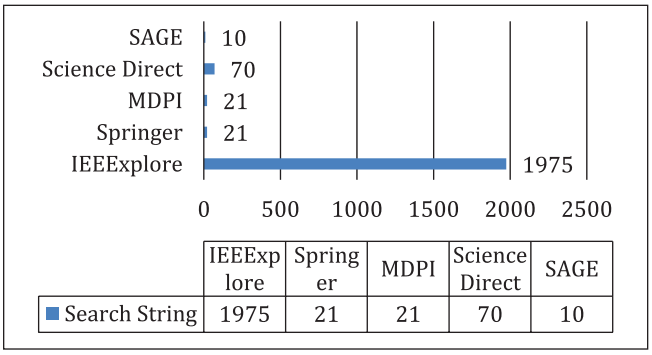

Figure 6. Chart indicating the number of records of each search string.

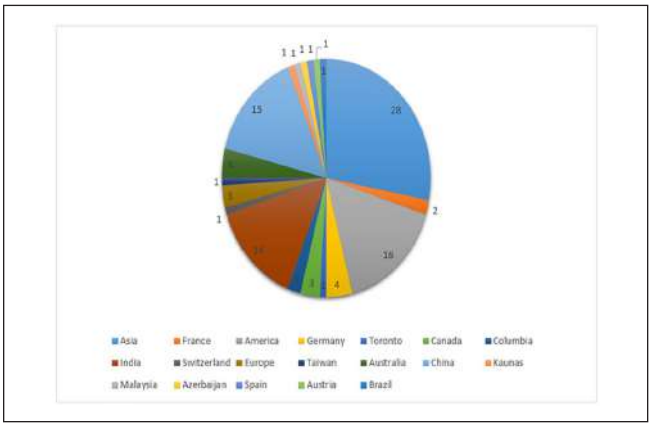

Figure 7. The systematic review's article distribution by country.

Inclusion Criteria

The search string was used as the inclusion criteria, and studies that fulfilled following describes the inclusion criterion for writing research paper. Table 9 provide the inclusion criteria for each question which were applied for the selection of review papers.

Exclusion Criteria

There is the following exclusion criterion for writing review paper is listed in Table 10.

Quality Assessment

Quality evaluation (QA) is common in systematic literature reviews, but it is less common in systematic mapping research. After reading our articles, we concentrated on assessing the research's relevance to our findings. We examined the breadth of each inquiry to determine if it fit with our objectives which was really useful in addressing our study query. Examine each item to ensure that it has clear instructions that show us precisely what we need to do. Table 11 clearly depicts the quality assessment criteria of selected research articles. These criteria are based on eight quality assessment questions (QAs).

Queries for Quality Assurance (QA) must assess the nature of the investigation of each assertion and provide a quantifiable correlation between each proposition.

The relevant measurement was performed according to the given indications to determine the evaluation linked to the impact factor of each of the articles (Table 12).

Each of the papers was rated on a scale of poor to exceptional for their major contributions to our systematic review, based on the degree of creativity, proposal details, validation, findings, and analysis, as well as references and number of citations. Table 12 displays the impact factor and rating for each item.

For the evaluation of metrics linked to bibliographic references, we established criteria in terms of the number of references and the proportion of references. We utilized Google Scholar analytics to determine the amount of citations. Table 13 depicts the overall score will range between 0 and 16, with values ranging from insufficient (0–2) to sufficient (3–5), good (6–10), very good (11–13), and outstanding (14–16).

Data Extraction and Analysis

The technique for data extraction was to provide a set of likely responses to research queries.

RQ1. What are the privacy and security issues of Big Data?

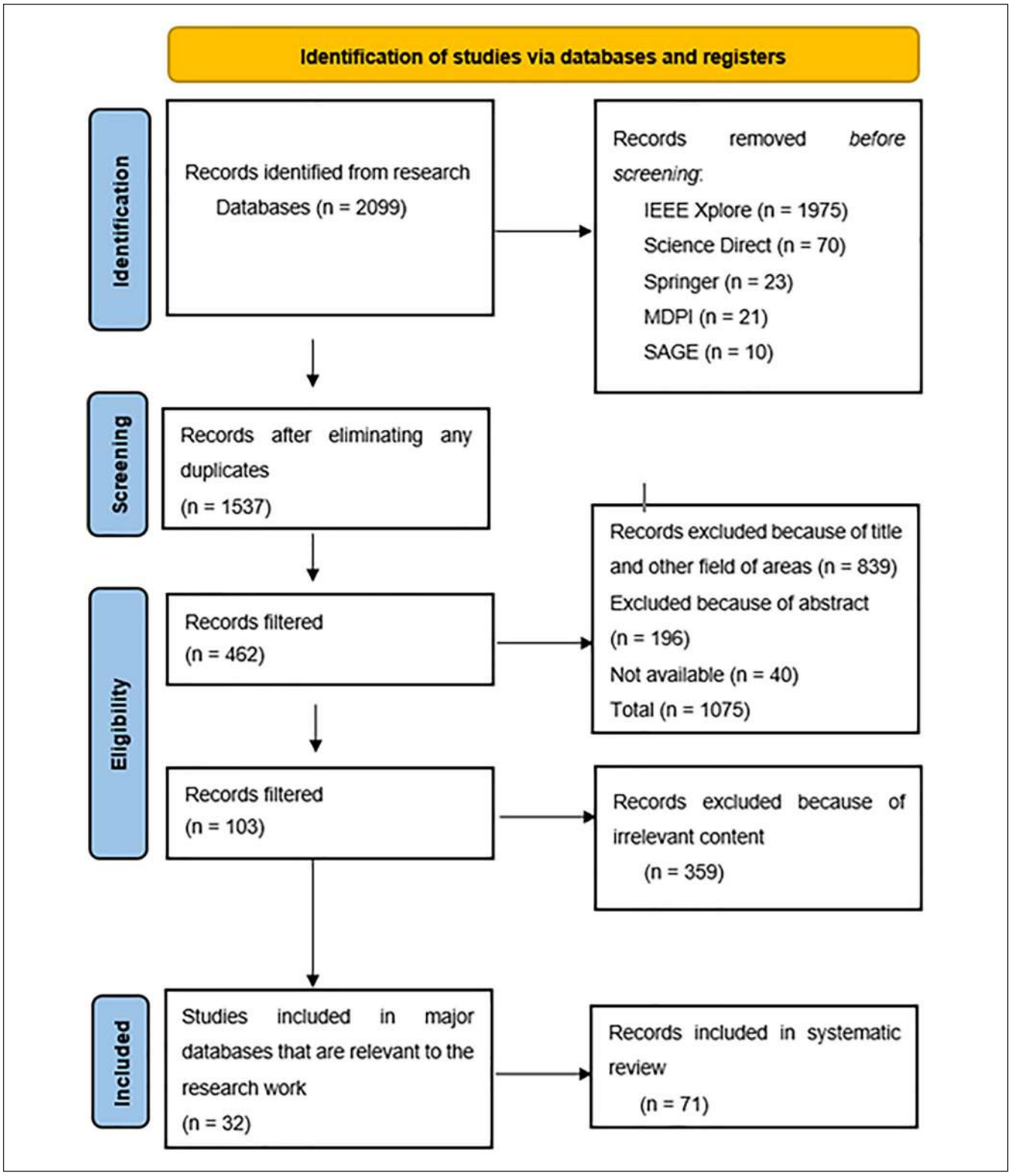

Figure 8. Flowchart depicting the study selection process.

To respond to this query, this article examines a wide range of Big Data safety and privacy analytics. Big Data varies from traditional technology in four aspects, according to current research: volume, diversity, speed, and value. The velocity, diversity, and volume of massive data present new security concerns including a large cloud infrastructure, the

Table 8. Inclusion and Exclusion Criteria.

Criteria	
Inclusion	Exclusion
Documents completed	Publications duplicated
Articles of the journal or conferences	Letters, reviews, and other publications to the publisher
The title and the abstract incorporate the search terms.	Publications that are not connected with Big Data or theme
	Not accessible publications

Table 9. Inclusion Criteria.

Sr.No	Inclusion criterion
1C1	All Published research papers that can answer the research questions
1C2	All published papers, journals, and books that are written in English language and according to Big Data applications and privacy prevention of Big Data analytics
1C3	Studies that were subjected to peer review
1C4	The studies which provide more knowledge and prove helpful for writing answer of my research questions
IC5	Studies that described the privacy prevention, security, and their impact on Big Data applications

Table 10. Exclusion Criteria.

Sr.No	Exclusion criterion
EC1	Papers that have not been published in English
EC2	Duplicate papers
EC3	Literature work which are not giving the clear idea of my research objective
EC4	Secondary data, such as magazines, articles, and reviews

Table 11. Quality Metrics.

	Metrics	Value	Weight
About the text of the article itself	M1: The abstract contains facts as well as a fair assessment of what was done and discovered	0/1	1
	M2: Give the qualifying criteria, as well as the sources and procedures of participant selection	0/1	1
	M3: Provides system architecture and component information	0/1	1
	M4: Details about the system's validation are provided	0/1	1
	M5: Results are provided based on objectives, limitations, and analysis	0/1	1
	M6: Provides a thorough and fair examination	0/1	1
	M7: References	0/1	1.5
	M8: (Conference/Journal) Type of publication	0/1	1

Table 12. Quality Assessment of the Selected Papers.

References	Ranking	Impact factor
Ur Rehman et al. (2016)	Q4	3
Koo et al. (2020)	Q2	3.2
Florea and Florea (2020)	Q2	3.2
D. Zhang (2018)	Q1	8
Varshney et al. (2020)	Q1	11
Guo et al. (2021)	Q2	2.7
Butpheng et al. (2020)	Q2	2.7
Awan, Khan, et al. (2021)	Q3	2.3
Technology (n.d.)	Q4	1
Sharif et al. (2021)	Q3	3.2
Shaukat et al. (2020)	Q3	3.0
Xiang et al. (2016)	Q3	3.2
Abouelmehdi et al. (2018)	Q1	11
Rajan et al. (2012)	Q4	1
Soria-Comas and Domingo-Ferrer (2016)	Q1	8
Sivan and Zukarnain (2021)	Q2	2.7

distinction between data source and design, and the cascading nature of data collection. These develop fast and force protected suppliers to adjust their response to threats and attacks (Rajan et al., 2012).

Cloud Computing provides three analytical and data storage services. The three departments are:

Storage

1. Software services platform
2. Infrastructure services
3. Infrastructure

Transfer

New protocols and approaches are required to deal with the difficulties of Big Data transfer. It's not enough of FTP and SECURE COPY (SCP). Two recent developments aiming at tackling massive problems with the flow are GRID FTP and GLOBUS.

Privacy and Security

It is one of our time's most serious challenges. Security is a state that is devoid of danger or harm whereas privacy is a state that others do not witness or disturb. However, many nowadays are worried that the quantity of scams, spam, dangerous URLs, and other hazards will increase their personal information on the Web. In terms of cybercrime, India ranks 11th among the top 19 nations (Soria-Comas & Domingo-Ferrer, 2016; Figure 9).

Following are some other important issues in data security:

The "privacy" is defined as a person or group's right and capacity to check their own access and use of information (Alier et al., 2021). In a variety of sectors, enormous

Table 13. Article Evaluation Using Quality Metrics.

References	M1	M2	M3	M4	M5	M6	M7	M8	Total
Ur Rehman et al. (2016)	0	0	1	0	0	0	1	1	3
Koo et al. (2020)	1	0	0	0	1	0	1	0	3.2
Florea and Florea (2020)	0	0	1	1	1	0	0	0	3.2
Chen and Lin (2014)	1	2	0	0	0	0	0	0	3.3
Gai et al. (2016)	0	0	0	0	1	0	1	1	3.3
D. Zhang (2018)	1	1	0	2	1	1	1	1	8
Craigen et al. (2014)	0	0	1	0	0	0	0	0	1
Varshney et al. (2020)	2	1	1	1	1	2	1	2	11
Guo et al. (2021)	0	0	1	0	0	0	0	1	2.7
Butpheng et al. (2020)	0	0	0	1	1	0	0	0	2.7
Awan, Khan, et al. (2021)	0	2	1	0	0	0	0	0	2.3
Technology (n.d.)	0	0	0	0	1	0	0	0	1
Sharif et al. (2021)	1	1	0	0	1	0	0	0	3.2
Shaukat et al. (2020)	1	1	0	0	1	0	0	0	3.0
Cheng et al. (2017)	1	1	2	0	0	0	0	0	3.3
Dev Mishra and Beer Singh (2017)	0	0	0	0	0	1	1	1	3.3
Q. Zhang et al. (2016)	1	2	0	0	0	0	0	0	3.3
Yu (2016)	0	0	0	0	0	1	1	1	3.3
Xiang et al. (2016)	0	2	1	0	0	0	0	0	3.2
Kantarcioglu and Shaon (2019)	0	0	0	0	0	1	1	1	3.3
Abouelmehdi et al. (2018)	2	1	2	1	1	1	1	2	11
Shamsi and Khojaye (2018)	0	0	0	0	0	1	1	1	3.3
Rajan et al. (2012)	0	0	0	0	0	1	0	0	1
Soria-Comas and Domingo-Ferrer (2016)	1	1	2	1	1	1	1	0	8
Singh et al. (2018)	1	0	0	0	0	0	0	1	2
Alguliyev and Imamverdiyev (2014)	0	0	0	0	1	1	0	1	3.3
Dev Mishra and Beer Singh (2017)	0	0	0	0	0	1	1	1	3.3
Sivan and Zukarnain (2021)	0	0	0	0	0	0	1	1	2.7
Cárdenas et al. (2013)	0	1	0	0	0	0	1	0	1.71
Alshboul et al. (2015)	0	0	1	0	1	1	0	0	3.3
Alam and Awan (2018)	0	0	1	1	1	0	0	0	3.3
Sweeney (2002)	1	0	0	0	0	0	0	0	1

volumes of data are created and analyzed every day. As a result, systematic processes may be used to ensure data privacy (Singh et al., 2018). Big Data analysis facilitates abuses of privacy. True, the impact of the privacy of end-users is yet to be fully understood. Big Data applications must be created that respect the privacy of people (Dumitras & Shou, 2011).

Employee privacy is threatened throughout the whole data life cycle, which may be based on privacy issues raised by data processing (Ebert et al., 2021).

Big Data analytics for APT detection. New algorithms for detection are needed to analyze enormous volumes of data from a range of sources. There are a few proof-of-concept installations with promising results utilizing Big Data analytics for security event detection (François et al., 2011).

High-performance encryption and decryption techniques; encrypting the search, encoding the attribute, and targeting the availability, reliability, and integrity of large-scale data (Yen et al., 2013).

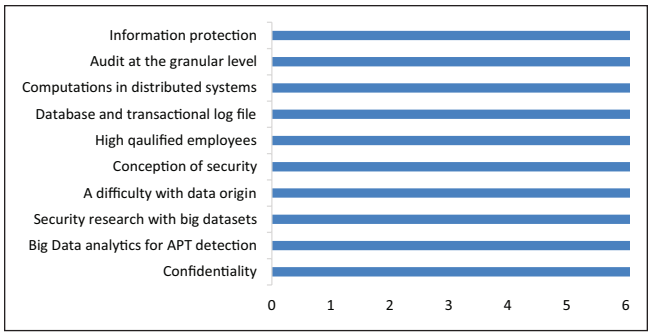

Figure 9. Some important privacy and security concerns of Big Data confidentiality.

Security research with Big Data datasets. A great deal of information is available, but it is virtually hard to separate the ground reality from organically generated data. These datasets may contain plenty of data, but it is difficult to

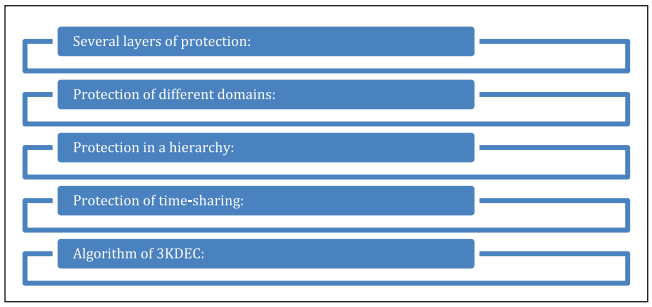

Figure 10. Possible solutions to protect user's privacy and security.

determine what is good and/or where attack data may be obtained (Geist & Reed, 2017).

A difficulty with data origin. Because large data enables us to increase the amount of data sources we utilize, it is hard to determine whether any data source can be confident enough to generate accurate results in our analytical algorithms (Joseph et al., 2019). We thus need to check the correctness and integrity of our tools' data. To identify and minimize the consequences of unlawfully added data, we can use approaches from adverse machine learning and robust statistics (Anam et al., 2021; Yang et al., 2019).

Conception of security. In visualization, people are exceptionally able to identify patterns in pictures. Although the technology of visualization is still in its infancy, study and development is growing (Pham & Dang, 2019).

While security solutions for open source and commercial data visualization are available, the safety visualization of data is still very essential; the Excel spreadsheets are conquered by pie charts, graphs, and pivot tables (L. Zhang et al., 2012).

Highly qualified employees. Suitable qualified staff are an essential part of a successful Big Data deployment for information security. One of the problems is the lack of such staff in this respect. Specific examples of competencies are expertise in data management, data analysis, and threat analysis (L. Zhang et al., 2012). As these skills are improbable to be found in one individual, organizations will need to establish teams of collaborators to obtain the greatest results in their Big Data operations (Alguliyev & Imamverdiyev, 2014).

The subject of how data is utilized for security purposes has resurfaced as a political issue (Aradau & Blanke, 2015).

Both the database and the transaction log file are password secured. Big Data management today requires automatic tiring for availability and scalability with the rapidly increased data base capability. Auto-tired systems do not track the location of the database, providing a new problem for safe database storage (Dev Mishra & Beer Singh, 2017).

Computations in distributed systems that are secure. Parallelism is used to process exceptionally large amounts of data in computations and physical storage. A excellent example is the MapReduce Framework. The two primary attack preventative methods are the protection of mappers and the defense of data in the presence of untrusted mappers (Application et al., 1999).

Audits at the granular level. The problems of security are employed for a number of reasons, including compliance, regulatory, and forensics. One instance of the attack is missed. Granular auditing is utilized when dealing with data objects almost certainly assigned (Lee et al., 2001).

Information protection. Data security has evolved into a significant data problem in and of itself. Dealing with large amounts of data is a difficult task from a security sense. Big Data is being used by data analysts to learn about our buying habits, health state, sleep cycles, movement patterns, online usage, friendships, and so on (Zwitter, 2014). A few examples are financial services, commercial websites, social networking sites, the health sector, networking, and anomaly detection. As a result, information security is one of the Big Data challenges. Customers' and workers' personal information should be kept private (Nair et al., 2016). According to the survey, 40 bogus accounts, 25 pages, 6 groups, and 28 Instagram accounts were banned on Pakistan by Facebook which in May tried to influence public opinion (Rastogi et al., 2018; Figure 10).

RQ2. What are the possible solutions/measures may be put in place to protect user's privacy and security?

To respond to this query, the possible solutions in order to protect user's privacy are as follows:

Several layers of protection. The security of an information system is established in computer hardware through layer expansion. The outer layers' security is reliant on the internal layers' security. The more coatings there are, the greater the security (Kuhn et al., 2005).

Protection of different domains. The four DNS categories comprise the local area, network border, network broadcast, and infrastructure. Consequently, numerous technologies are used in different approaches to protect areas to develop a distributed security system (Stouffer et al., 2011).

Protection in a hierarchy. Because the value of the same information varies depending on the institute. As a result, hierarchical security is required, and in this situation, various

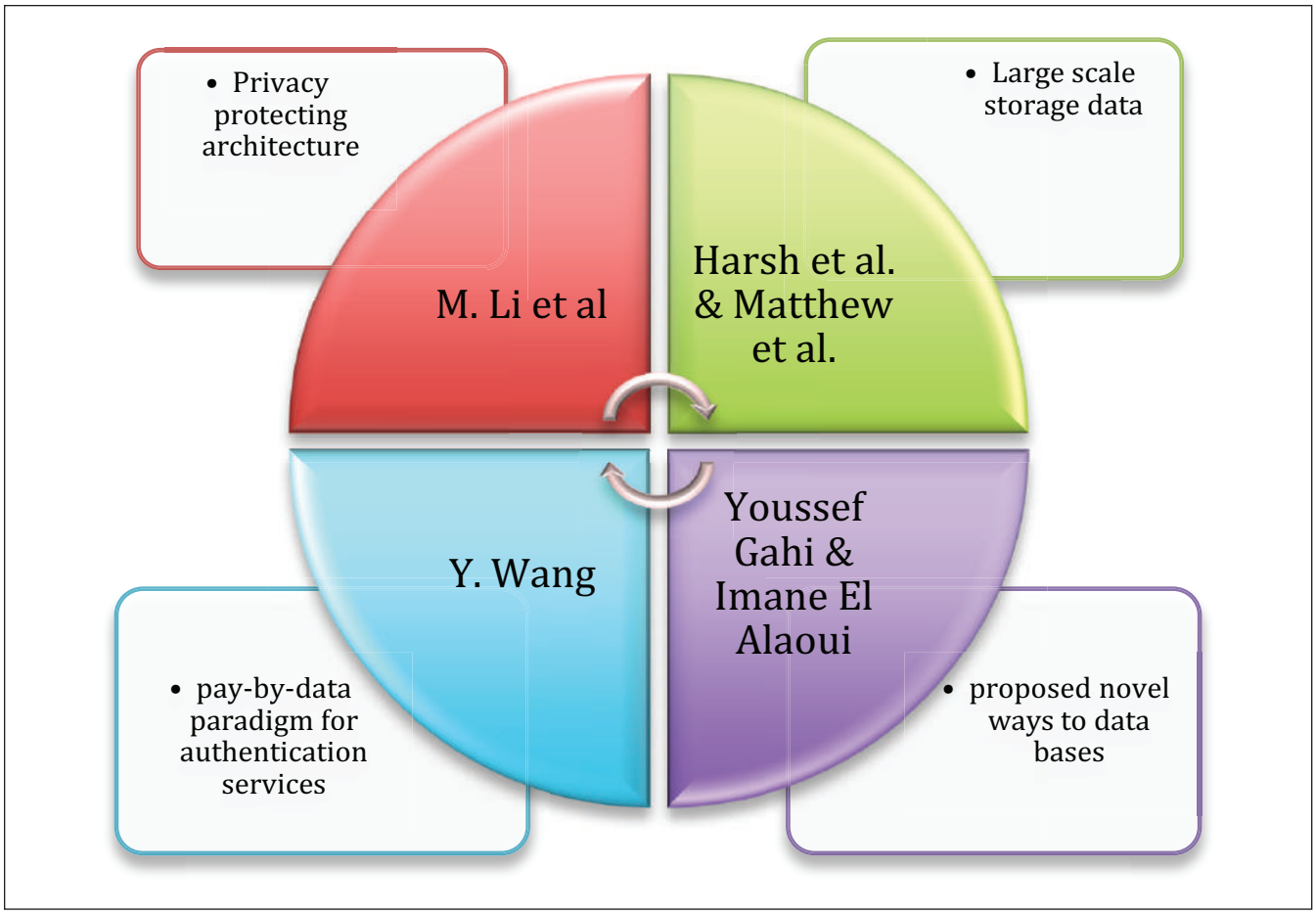

Figure 11. Types of procedures to preserve consumer's privacy.

access control techniques are employed so that a single user is responsible for just one specific parameter (Zissis & Lekkas, 2012).

Protection of time-sharing. Data security is a fluid process with Big Data. Taking the time to think about Big Data security may greatly improve your results (Inbarani & Kumar, 2015).

Privacy is the key needs for users data exchange or access (Sivan & Zukarnain, 2021).

Algorithm of 3KDEC. To provide a feasible solution to the problem of transforming numeric data to alphanumeric type and therefore encrypted data not being stored in existing numeric fields, a symmetric key block encryption approach is used (Figure 11).

The types of procedures that might be taken to preserve consumers' privacy are as follows:

The existence of legislation that protects user privacy is the most critical condition for increasing user privacy. We classified likely data privacy issues in Big Data systems into four categories based on a literature review: Tracking by the government, data gathering by service providers, re-identification attacks, and data violations are all examples of data violations (International Standard Organization, 2011). The main objectives of data theft were financial systems and public databases. Since theft of data from such systems might have the most devastating consequences, this is appropriate (Verizon, 2016).

Many businesses have quickly implemented large-scale data analysis, before ensuring that all elements, notably security and privacy, are taken into account. In striving to rebuild and enhance the core idea for Big Data analysis, the scientific community has given significant attention to this expanding field. The proposed solutions of securing Big Data analytics are summarized in Table 14.

Our key function in this research is to provide unique ways for protecting data secrecy (Jacobs & Popma, 2019).

The most crucial requirement to improve privacy of users is the provision of legislation that can protect user confidentiality. Then measures are employed to secure privacy and data anonymity. To avoid data theft, physical and security precautions must be taken. These solutions are related. It should be noted. Strong standards, for example, might

Table 14. Proposed Solutions of Securing Big Data Analytics.

Article	Proposed solution	Studies
M. Li et al.	There is a privacy-protecting architecture in the cloud and Big Data scenarios. The recommended design minimizes node authority by prioritizing them to avoid cloud providers from inspecting and tracking the activity of users. The new architecture also makes it possible for Cloud clients to personalize their privacy protection while restricting the supplier's capacity to modify confidentiality.	Li et al. (2013)
Harsh et al. and Matthew et al.	The authors examined the need for large-scale storage and noted many possible challenges for huge data. Harsh et al. and Matthew et al. highlighted Big Data problems, respectively, for medical and social media.	Cárdenas et al. (2013)
Y. Wang	The authors developed a pay-by-data paradigm for authentication services to ensure access to data generated by users.	Alshboul et al. (2015)
Youssef Gahi and Imane El Alaoui	The authors have proposed novel ways to data bases, employing Big Database Analytics rather than data processing protection.	Gahi and Alaoui (2019)
Lei Xu and Xiaoxin Wu	CL-PRE is the author-defined way of certificate-less proxy encryption. This bases on random and re-encoding keys and allows a data owner to share its data while limiting access control and decreasing the confidence in cloud infrastructure.	Xu et al. (2012)

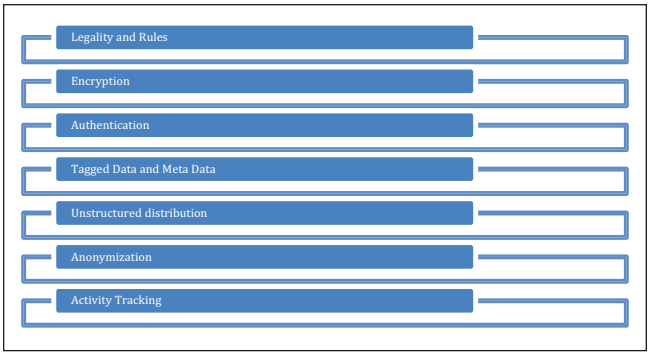

Figure 12. Possible techniques of Big Data.

compel service providers to take action in order to avoid data infringement. In addition, legislation might require service providers to adopt data protection through design which includes data protection safeguards in the design and development process (Colesky et al., 2016).

Figure 12 shows the possible techniques in command to defend confidentiality of Big Data are as follows:

The Big Data Platform consists of various technical developments, for both storage and processing. The effective and direct application of the traditional security approaches to the large data situations is not possible. They offer a collection of methods which may be utilized to secure large-scale data.

Legality and rules. Big Data is an enormous phenomenon that continues to transform the globe. There are insufficient laws and regulations governing Big Data mining. Big Data might include financial, health, and personal sensitive information. Choose the site of storage and processing to comply with the agreements of nations is crucial.

Encryption. Big Data may be utilized for storage, calculations, and communications security. The fact that data may be collected in bulk and kept for later analysis may have a "chilling effect" on society (McDermott, 2017). If hostile or unauthorized nodes have access to the storage, no information can be extracted from the storage as only parties with decryption keys may peek inside. Blind processing techniques are an appropriate option based on homomorphic encryption.

Authentication. Authentication techniques are a good approach to keep track of who has access to what. Such a strategy should be used in the design of Big Data solutions to manage both cluster membership and access to key storages.

Tagged data and meta data. The use of metadata and tag techniques is another option for classifying the gathered data according to its relevance and whether or not it should be included in the treatment. In this approach, Big Data technologies will not treat all records the same and it will be able to handle private data.

Distribution, that is, unstructured. It would be interesting to find a technique for making it impossible for parties without access to the global system to use data if a mischievous party were to have admittance to one or more clusters. There are numerous data extraction strategies, as well as distinct data kinds (Alam & Awan, 2018).

Big Data's distributed mode might be precisely tailored to meet this goal. To do this, we must avoid keeping correlated data in the same cluster and instead disperse it over multiple. Furthermore, unstructured distribution allows data to be separated from individual associated information, preventing hackers from obtaining relevant information even if they have access to some nodes.

Anonymization. Anonymization is another way for safeguarding collected Big Data. The basic aim is to safeguard the association of persons with vital information by utilizing

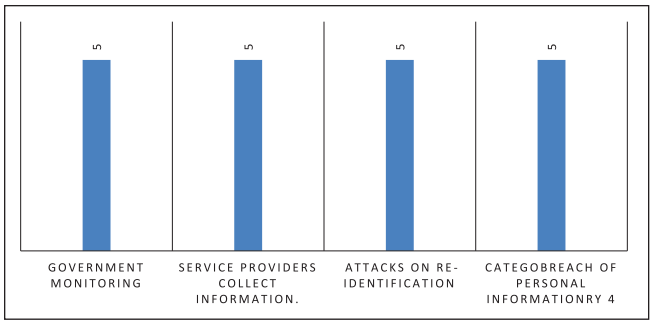

Figure 13. Forms of privacy violations in Big Data.

data perturbation and data swapping techniques. There's also *k*-anonymity, which prevents data from being re-identified by masking its true location among $k-1$ others (Sweeney, 2002). *K*-anonymity can prevent identification information, which makes it hard to link the *k*-anonymized record accurately to the original data set (Csányi et al., 2021).

Activity tracking. It is necessary to log all activities conducted over large data, as well as the individuals accountable for these acts, in order to safeguard or at least oversee the data. These logs might be examined to see if there were any attempts to modify the large data that were harmful.

RQ3. What are the different forms of privacy violations that arise in Big Data systems?

To respond to this query, the authors give an introduction of data privacy and define four kinds of privacy abuses on Big Data systems as well as analyze the strengths and limitations of their protection measures in order to address this issue. They also offer tips on how to protect your privacy. Figure 13 shows the types of privacy violations in Big Data analytics.

The forms of confidentiality violations in Big Data systems are as follows:

Based on a study of the literature, we classified potential data privacy violations in Big Data systems into four categories: government tracking, data collection by service providers, re-identification attacks, and data transgressions. A collection of various kinds of privacy infringements as well as samples and actual life instances are given in Table 15.

Figure 14 illustrates an overview of frequently used platform for privacy violations as shown below.

RQ4. What are encryption techniques for various activities in Big Data analytics?

To respond to this query, the main focus of the paper is on preserving the privacy of data using anonymization methods (Liu et al., 2018). Every algorithm has its own set of advantages and disadvantages and each algorithm responds in different ways (Mohan et al., 2014). Few types of encryption algorithms used in Big Data analytics are summarized in Table 16.

Figure 15 shows the types of anonymization techniques used in Big Data analytics.

RQ5. What are the suggestions for restricting privacy violations?

To respond this query, the suggestions for restricting privacy violations are as follows:

To address this issue, only a few methods are available to prevent governments and service companies from invading the privacy of the public. Such privacy breaches appear to benefit the Service and users, in general, are ready to abandon their privacy to assist them. Re-identification and data breaches, on the other hand, can be tremendously destructive (Pötzsch, 2009). Figure 16 depicts the suggested model for minimizing privacy violations, illustrating four distinct forms of data protection violations, and the roles of various stakeholders in mitigating their effects.

There is more collaboration in the double-sided arrows between objects whereas single sided arrows highlight the necessity for a specific approach to prevent a certain kind of protection offence.

In a more general sense, the public is not aware of the challenges of privacy. The main preconditions for preventing abuses of privacy are public awareness and highlighting measures to limit access to personally identifiable information. The usage of feedback awareness tools can do this. For example, in social networks, where friends can make inappropriate images, such tactics may be helpful (Pötzsch, 2009). Confidential gathering by service providers and governments may be minimized through user-friendly policies. Legislation should be designed to ensure privacy instead of privacy of choice. Laws are essential to civil society and governance and the promotion of awareness. The spreading of social ethics to protect privacy can also be helped by civil society. In order to securely keep personal information, several conditions must be followed. If there is a data breach, service providers must inform victims as quickly as feasible. In order for companies to understand the necessity to secure personal data, the governments need likewise implement legislation.

Informaticians also play a key role. They must enhance the safety of mass storage technologies. To achieve this, cyber defense systems must be strengthened and efficient encryption and storage solutions should be provided. Interruption detection and prevention should be upgraded to lower the occurrence of data violations. Data access procedures should also be standardized in order to assure permission. The variability in storage methods for huge data systems complicates matters. The research community should also work on the usage of semantic web techniques like data origin and linkages to develop individual data protection rules and detect unintended data inferences (Farkas, 2014). Anonymization processes should be enhanced in order to diminish the effect of re-identification.

Table 15. Types of Privacy Violations.

Form	Description	Examples/real-life incidents	Studies
Government monitoring	Governments employ monitoring programs to enhance security. Sensitive information may be obtained in a number of ways.	PRISM: The United States administration collects data from key intelligence providers. Monitoring: Cities gather data to enhance services such as traffic surveillance.	Shamsi and Khojaye (2018)
Service providers collect information.	A service provider is able to collect and utilize personal information from a user. It must be remembered that privacy might be unintentionally violated.	Auto-scan: Significant ads are shown via e-mails or postings on websites of the social media. Google documents shared with other people accidentally: Google unintentionally shared user papers with others.	Khan et al. (2019)
Attacks on re-identification	The combination of Big Data sets may distinguish individuals.	Medical insurance and elector registration records were connected to identify sensitive governor information.	De Goede (2014)
Breach of personal information	A source of personal data can be hacked and displayed.	Ashley Madison: A dating website has been hacked and has been exposed. Personal information on 157,000 UK's major telecommunications provider users has been exposed. Experian says hackers obtained sensitive information such social refuge and ID numbers. Target: Information on private credit cards was gathered via Target's point-of-sale terminals.	Cheng et al. (2017)

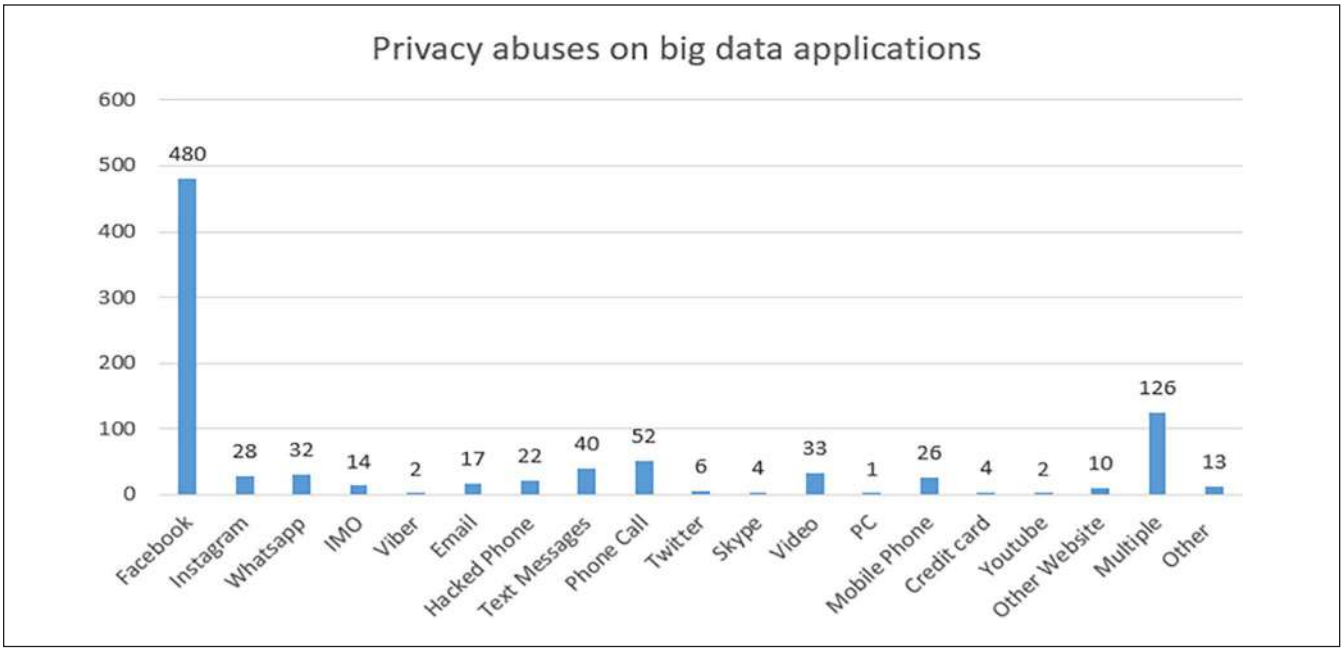

Figure 14. Frequent platform for privacy violations.

Finally, to foster innovation and standardization, governments should help R&D.

Discussion

Big Data is a relatively new IT concept, and it is apparent that further research is needed in this area. Many papers, however, show a substantial gap, suggesting that research is skewed toward traditional techniques and that Big Data is under-researched. In the bulk of the articles, the study findings are one-sided and incomplete. This drawback applies to all of the articles utilized, with the exception of article (Dev Mishra & Beer Singh, 2017), which covers a wide range of topics and solutions for security concerns.

Both perspectives are presented in Table 17. This occurs mostly in e-Commerce scenarios as example of a typical method oriented to customer service providers and, secondly, in web forums that demonstrate how people engage arbitrarily.

Scholars can examine issues like social networks, multimedia, commercial organizations, companies, the business environment, anonymity protection, data watermarking, data

Table 16. Types of Encryption Algorithms Used in Big Data Analytics.

Scheme of anonymity	Description	Weakness/attack	Studies
K-anonymity	There are at least k redundant identifications (QIDs), to provide anonymity for k1 people.	If all documents contain sensitive information, their confidentiality might be compromised. Hintergrund Attack: Employing profound evidence can be identified by using background knowledge.	Ninghui et al. (2007)
L-diversity	At least one "well represented" value in the distribution of a sensitive characteristic applies to each equivalence class.	A similarity attack can be used by an adversary to detect possible sensitive information. Skewness attack: In some parts of the data sensitive data may be recognized as it differs greatly from the distribution of sensitive data throughout the remaining dataset.	Ninghui et al. (2007)
T-closeness	The frequency distribution between sensitive features should be "close" in each equivalent class to the circulation of sensitive characteristics throughout the data set, where *t* is the predefined threshold value.	There are no computational solutions for *t*-closeness while decreasing data loss. Data loss is therefore feasible when searching for *t*-closeness.	Soria-Comas and Domingo-Ferrert (2013)
Differential privacy	The aim is to restrict sensitive data disclosure by reducing individuals' impact during the survey. This is achieved by introducing sounds to the selected results (e.g., the Laplace and geometric mechanisms).	Infringements of data may arise due to improper disclosure of original data. This approach, for example, maintains the privacy of people by making the query results sufficient noise; nonetheless, the data remains on the server where data infringements occur.	Soria-Comas and Domingo-Ferrert (2013)

Provence, role-based access control, and risk-adaptive access control. In most articles, a larger variety of facts must be studied in order to reach more definite conclusions. Furthermore, there is a lack of comparison of alternative options, as well as why or how they may be applied, in these papers. The material included in this study also includes gaps in terms of how specific security measures may be implemented. Second, most papers lack technical information on solutions, such as methods used to solve specific issues; case studies, for example, might be beneficial in learning and getting answers in a certain sector. Third, many publications discuss security problems in a broad sense without providing specific information about a specific problem in a given region. Most frequent security and privacy concerns, in particular, lack specifics on what privacy entails and what information must be secured. Technological advancements can offer a number of benefits, but they can also present risks that might posture a risk and result in a violation of privacy. If corporations make sensitive information public, they may face substantial fines. Big Data is a relatively new concept that refers to the massive quantity of data that must be processed and stored in directive to prevent data breaches. In many sectors of Big Data, there are several security concerns, and sensitive data must be safeguarded. Though, because this field is still new, a lot of study in Big Data is needed, and there are a lot of issues that need to be answered, there is a lot of research that needs to be done. The findings clearly show that security problems are identical across domains, that solutions may be similar across areas, and that many solutions rely on encryption methods. Furthermore, access to large amounts of data must be protected. Overall, this work has contributed to the body of knowledge about massive data security problems and highlighted research gaps in the field. On the contrary, this is an under-researched issue that needs to be thoroughly investigated.

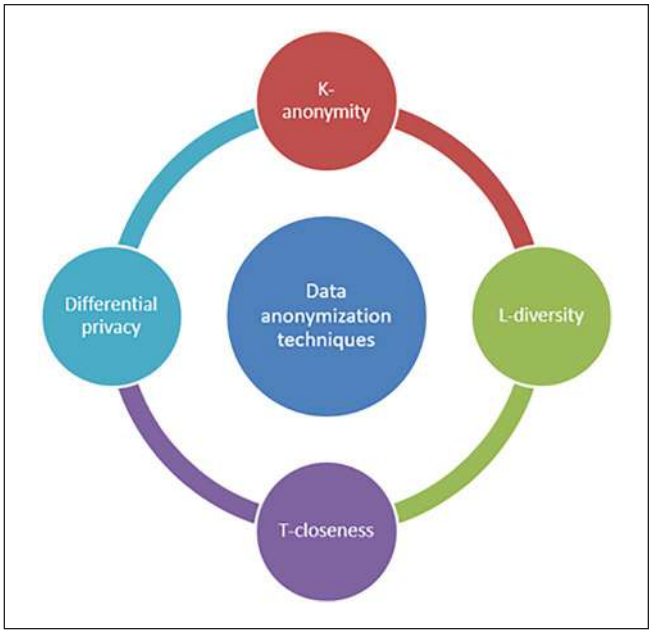

Figure 15. Types of anonymization techniques.

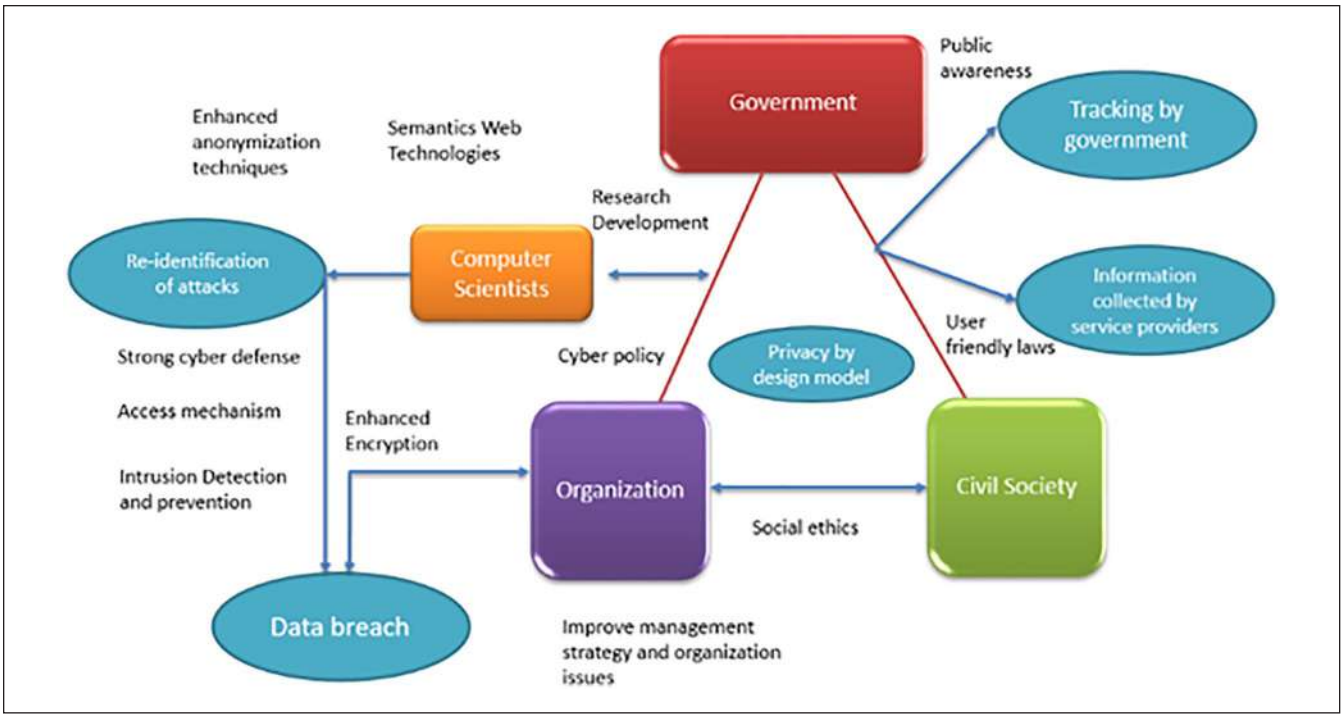

Figure 16. Shows a model for enhancing privacy.

Conclusion

Many companies and governments in many industries are using Big Data as a foundation to automate the processing and extraction of important insights to aid decision-making. The fact that all conceivable and diverse data must be collected and computed may, however, result in numerous security and privacy breaches. Although Big Data systems have proven beneficial for analysis and forecasting, it is important to implement protocols for keeping and preserving confidential data on Big Data platforms. A substantive and organized effort is required to achieve this crucial goal. The Six V's notion of Big Data were viewed as important issues. In order to build a more secure data processing and data computing infrastructure for the study, we intend to highlight several challenges about security and privacy. As there are so many safety data, security has become a huge data challenge. If we ignore or cannot examine any of these facts, Big Data security analytics technologies are not safe. The aim of Big Data analytics for safety is to obtain information that can be activated in real time. While Big Data analytics have a lot of promise, they still have a way to achieve full potential. Numerous security procedures were submitted for Big Data Analytics. A particular protocol should be used because of the safety issues and the application. Large data analytics focus on security and privacy issues and enhance the safety and privacy of Big Data platforms.

Table 17. Pros and Cons of Disclosure of Big Data (Pötzsch, 2009).

Applications	Pros	Cons
Web collectives	Social trade Relationships Collaborations Reputation	Theft of Identity spam marketing Stalking, enlistment In other settings, negative reputations
e-Commerce	Convenience Processing automated Price bonus Chosen information	Discrimination on prices spam marketing Theft of Identity

We've identified a number of security and privacy issues that Big Data technologies should consider in this article. We also discussed some potential solutions and strategies for protecting this distributed system. We also address few privacy violations and also discussed encryption algorithms used in Big Data analytics. Some of these protective measures will be included in an open source Big Data analysis tool as part of future development. Currently, several privacy-preserving strategies for Big Data exist, such as anonymization protection technology, access control, encryption, unstructured distribution, data tracing, differential privacy protection, anonymization, and so on. We end with a few recommendations for improving the efficiency of a Big Data project, and provide secure possible techniques and proposed solutions and model that minimizes privacy violations,

showing four different types of data protection violations and the involvement of different entities in reducing their impacts. However, in algorithms as well as in system areas, further research is needed to deal with the increasingly many problems ahead.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Haitham Nobanee https://orcid.org/0000-0003-4424-5600

References

Abdulhamid, S. M., Abd Latiff, M. S., Chiroma, H., Osho, O., Abdul-Salaam, G., Abubakar, A. I., & Herawan, T. (2017). A review on mobile SMS spam filtering techniques. *IEEE Access*, *5*, 15650–15666. https://doi.org/10.1109/ACCESS.2017.2666785

Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: Preserving security and privacy. *Journal of Big Data*, *5*(1), 1–18. https://doi.org/10.1186/s40537-017-0110-7

Adjei, J. K., Adams, S., Mensah, I. K., Tobbin, P. E., & Odei-Appiah, S. (2020). Digital identity management on social media: Exploring the factors that influence personal information disclosure on social media. *Sustainability (Switzerland)*, *12*(23), 1–17. https://doi.org/10.3390/su12239994

Aftab, M. O., Javed Awan, M., Khalid, S., Javed, R., & Shabir, H. (2021). *Executing spark BigDL for leukemia detection from microscopic images using transfer learning* [Conference session]. 2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021, Riyadh, Saudi Arabia, pp. 216–220. https://doi.org/10.1109/CAIDA51941.2021.9425264

Ahmed, H. M., Awan, M. J., Khan, N. S., Yasin, A., & Shehzad, H. M. F. (2021). Sentiment analysis of online food reviews using Big Data analytics. *Ilkogretim Online*, *20*(2), 827–836. https://doi.org/10.17051/ilkonline.2021.02.93

Alam, T. M., & Awan, M. J. (2018). Domain analysis of information extraction techniques. *International Journal of Multidisciplinary Sciences and Engineering*, *9*(6), 1–9.

Alguliyev, R., & Imamverdiyev, Y. (2014). *Big Data: Big promises for information security* [Conference session]. 8th IEEE International Conference on Application of Information and Communication Technologies, AICT 2014 – Conference Proceedings, Astana, Kazakhstan. https://doi.org/10.1109/ICAICT.2014.7035946

Ali, Y., Farooq, A., Alam, T. M., Farooq, M. S., Awan, M. J., & Baig, T. I. (2019). Detection of schistosomiasis factors using association rule mining. *IEEE Access*, *7*, 186108–186114. https://doi.org/10.1109/ACCESS.2019.2956020

Alier, M., Jose, M., Guerrero, C., Amo, D., Severance, C., & Fonseca, D. (2021). *Privacy and E-learning : A pending task. Sustainability*, *13*(16), 9206.

Alshboul, Y., Wang, Y., & Nepali, R. K. (2015). *Big Data lifecycle: Threats and security model* [Conference session]. 2015 Americas Conference on Information Systems, AMCIS 2015, Fajardo, Puerto Rico, pp. 1–7.

Ambalavanan, V. (2020). Cyber threats detection and mitigation using machine learning. In P. Ganapathi & D. Shanmugapriya (Eds.), *Handbook of research on machine and deep learning applications for cyber security* (pp. 132–149). IGI Global.

Anam, M., Ponnusamy, V., Hussain, M., Nadeem, M. W., Javed, M., Goh, H. G., & Qadeer, S. (2021). Osteoporosis prediction for trabecular bone using machine learning: A review. *Computers, Materials and Continua*, *67*(1), 89–105. https://doi.org/10.32604/cmc.2021.013159

Application, F., Data, P., Examiner, P., & Andrews, M. (1999). *United States Patent (19)*. United States Patent.

Applications, C., Technology, I., Engineering, S., Engineering, S., & Engineering, C. (n.d.). *Efficient Residential Load Forecasting using Deep Learning Approach Rida Mubashar * Mazhar Javed Awan Muhammad Ahsan Awais Yasin Vishwa Pratab Singh. X*(2006). United States Patent.

Aradau, C., & Blanke, T. (2015). The (Big) data-security assemblage: Knowledge and critique. *Big Data and Society*, *2*(2), 1–12. https://doi.org/10.1177/2053951715609066

Awan, M. J. (2020). Fake news classification bimodal using convolutional neural network and long short-term memory. *Article in International Journal of Emerging Technologies in Learning (IJET)*, *11*(5), 209–212.

Awan, M. J., Khan, M. A., Ansari, Z. K., Yasin, A., & Shehzad, H. M. F. (forthcoming). Fake profile recognition using big data analytics in social media platforms. *International Journal of Computer Applications in Technology*. https://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijcat

Awan, M. J., Khan, R. A., Nobanee, H., Yasin, A., Anwar, S. M., Naseem, U., & Singh, V. P. (2021). A recommendation engine for predicting movie ratings using a Big Data approach. *Electronics (Switzerland)*, *10*(10), 1215. https://doi.org/10.3390/electronics10101215

Awan, M. J., Rahim, M. S. M., Nobanee, H., Munawar, A., Yasin, A., & Zain, A. M. (2021). Social media and stock market prediction: A Big Data approach. *Computers, Materials and Continua*, *67*(2), 2569–2583. https://doi.org/10.32604/cmc.2021.014253

Awan, M. J., Rahim, M. S. M., Nobanee, H., Yasin, A., Khalaf, O. I., & Ishfaq, U. (2021). A Big Data approach to black Friday sales. *Intelligent Automation and Soft Computing*, *27*(3), 785–797. https://doi.org/10.32604/iasc.2021.014216

Awan, M. J., Rahim, M. S. M., Salim, N., Ismail, A. W., & Shabbin, H. (2019). Acceleration of knee MRI cancellous bone classification on google colaboratory using convolutional neural network. *International Journal of Advanced Trends in Computer Science and Engineering, 8*(1.6 Special Issue), 83–88. https://doi.org/10.30534/ijatcse/2019/1381.62019

Awan, M. J., Rahim, M. S. M., Salim, N., Mohammed, M. A., Garcia-Zapirain, B., & Abdulkareem, K. H. (2021). Efficient detection of knee anterior cruciate ligament from magnetic resonance imaging using deep learning approach. *Diagnostics*, *11*(1), 105. https://doi.org/10.3390/diagnostics11010105

Awan, M. J., Raza, A., Yasin, A., Muhammad, H., & Shehzad, F. (2021). The customized convolutional neural network of face

emotion expression classification. *Annals of R.S.C.B*, *25*(6), 5296–5304.

Barth-Jones, D. C. (2012). The "Re-Identification" of governor William Weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now. https://doi.org/10.2139/ssrn.2076397

Battams, K. (2015). *Stream mining for solar physics: Applications and implications for big solar data* [Conference Session]. Proceedings – 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, Washington, DC, pp. 18–26. https://doi.org/10.1109/BigData.2014.7004400

Butpheng, C., Yeh, K. H., & Xiong, H. (2020). Security and privacy in IoT-cloud-based e-health systems-A comprehensive review. *Symmetry*, *12*(7), 1–35. https://doi.org/10.3390/sym12071191

Cárdenas, A. A., Manadhata, P. K., & Rajan, S. P. (2013). Big data analytics for security. *IEEE Security & Privacy*, *11*(6), 74–76.

Chandramouli, B., Goldstein, J., & Duan, S. (2012). *Temporal analytics on Big Data for web advertising* [Conference Session]. Proceedings – International Conference on Data Engineering, Arlington, VA, pp. 90–101. https://doi.org/10.1109/ICDE.2012.55

Chandrasekar, Dr. C. (2018). Classification techniques using spam filtering email. *International Journal of Advanced Research in Computer Science*, *9*(2), 402–410. https://doi.org/10.26483/ijarcs.v9i2.5571

Che, D., Safran, M., & Peng, Z. (2013). From Big Data to Big Data mining: Challenges, issues, and opportunities. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7827 LNCS*, pp. 1–15. https://doi.org/10.1007/978-3-642-40270-8_1

Chen, X. W., & Lin, X. (2014). Big Data deep learning: Challenges and perspectives. *IEEE Access*, *2*, 514–525. https://doi.org/10.1109/ACCESS.2014.2325029

Cheng, L., Liu, F., & Yao, D. D. (2017). Enterprise data breach: Causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(5), 1–14. https://doi.org/10.1002/widm.1211

Colesky, M., Hoepman, J. H., & Hillen, C. (2016). A critical analysis of privacy design strategies. Proceedings – 2016 IEEE Symposium on Security and Privacy Workshops, SPW 2016, San Jose, CA, pp. 33–40. https://doi.org/10.1109/SPW.2016.23

Craigen, D., Diakun-Thibault, N., & Purse, R. (2014). Defining cybersecurity. *Technology Innovation Management Review*, *4*(10), 13–21. https://doi.org/10.22215/timreview835

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., & Orosz, T. (2021). Challenges and open problems of legal document anonymization. *Symmetry*, *13*(8), 1–25. https://doi.org/10.3390/sym13081490

De Goede, M. (2014). The politics of privacy in the age of preemptive security. *International Political Sociology*, *8*(1), 100–104. https://doi.org/10.1111/ips.12042

Dev Mishra, A., & Beer Singh, Y. (2017). *Big Data analytics for security and privacy challenges* [Conference session]. Proceeding – IEEE International Conference on Computing, Communication and Automation, ICCCA 2016, Greater Noida, India, pp. 50–53. https://doi.org/10.1109/CCAA.2016.7813688

Dumitras, T., & Shou, D. (2011). *Toward a standard benchmark for computer security research: The worldwide intelligence network environment (WINE)* [Conference session]. Proceedings of the 1st Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, Salzburg, Austria, BADGERS 2011, pp. 89–96. https://doi.org/10.1145/1978672.1978683

Ebert, I., Wildhaber, I., & Adams-Prassl, J. (2021). Big Data in the workplace: Privacy due diligence as a human rights-based approach to employee privacy protection. *Big Data and Society*, *8*(1). https://doi.org/10.1177/20539517211013051

Farkas, C. (2014). *Big Data analytics: Privacy protection using semantic web technologies* [Conference session]. NSF Workshop on Big Data Security and Privacy, Texas, San Antonio, United States.

Firdausi, I., Lim, C., Erwin, A., & Nugroho, A. S. (2010). *Analysis of machine learning techniques used in behavior-based malware detection* [Conference session]. Proceedings – 2010 2nd International Conference on Advances in Computing, Control and Telecommunication Technologies, ACT 2010, Jakarta, Indonesia, pp. 201–203. https://doi.org/10.1109/ACT.2010.33

Florea, D., & Florea, S. (2020). Big Data and the ethical implications of data privacy in higher education research. *Sustainability (Switzerland)*, *12*(20), 1–11. https://doi.org/10.3390/su12208744

François, J., Wang, S., Bronzi, W., State, R., & Engel, T. (2011). *BotCloud: Detecting botnets using MapReduce* [Conference session]. 2011 IEEE International Workshop on Information Forensics and Security, WIFS 2011, Iguacu Falls, Brazil. https://doi.org/10.1109/WIFS.2011.6123125

Gahi, Y., & Alaoui, I. El. (2019). A secure multi-user database-as-a-service approach for cloud computing privacy. *Procedia Computer Science*, *160*, 811–818. https://doi.org/10.1016/j.procs.2019.11.006

Gai, K., Qiu, M., & Zhao, H. (2016). *Security-aware efficient mass distributed storage approach for cloud systems in Big Data* [Conference session]. Proceedings – 2nd IEEE International Conference on Big Data Security on Cloud, IEEE BigDataSecurity 2016, 2nd IEEE International Conference on High Performance and Smart Computing, IEEE HPSC 2016 and IEEE International Conference on Intelligent Data and S, New York, NY, pp. 140–145. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.68

Geist, A., & Reed, D. A. (2017). A survey of high-performance computing scaling challenges. *The International Journal of High Performance Computing Applications*, *31*(1), 104–113. https://doi.org/10.1177/1094342015597083

Guo, J., Yang, M., & Wan, B. (2021). A practical privacy-preserving publishing mechanism based on personalized k-anonymity and temporal differential privacy for wearable iot applications. *Symmetry*, *13*(6), 1043. https://doi.org/10.3390/sym13061043

Gupta, M., Jain, R., Arora, S., Gupta, A., Awan, M. J., Chaudhary, G., & Nobanee, H. (2021). AI-enabled COVID-19 outbreak analysis and prediction: Indian states vs. union territories. *Computers, Materials and Continua*, *67*(1), 933–950. https://doi.org/10.32604/cmc.2021.014221

Gurajala, S., White, J. S., Hudson, B., & Matthews, J. N. (2015, July). Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach [Conference session]. Proceedings of the 2015 International Conference on Social Media & Society, torpnto, ON, Canada.

Inbarani, H. H., & Kumar, S. S. (2015). *Big Data in complex systems* (*Vol. 9*). Springer. https://doi.org/10.1007/978-3-319-11056-1

International Standard Organization. (2011). *International standard ISO/IEC information technology—Security techniques—Application security*.

Jacobs, B., & Popma, J. (2019). Medical research, Big Data and the need for privacy by design. *Big Data and Society*, *6*(1), 1–5. https://doi.org/10.1177/2053951718824352

Javed, R., Saba, T., Humdullah, S., Mohd Jamail, N. S., & Javed Awan, M. (2021). *An efficient pattern recognition based method for drug-drug interaction diagnosis* [Conference session]. 2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021, Riyadh, Saudi Arabia, pp. 221–226. https://doi.org/10.1109/CAIDA51941.2021.9425062

Joseph, A. D., Nelson, B., Nelson, B., & Tygar, J. D. (2019). *Adversarial Machine Learning*. Cambridge University Press. https://doi.org/10.1017/9781107338548

Jusas, V., Japertas, S., Baksys, T., & Bhandari, S. (2019). Logical filter approach for early stage cyber-attack detection. *Computer Science and Information Systems*, *16*(2), 491–514. https://doi.org/10.2298/CSIS190122008J

Jusas, V., & Samuvel, S. G. (2019). Classification of motor imagery using combination of feature extraction and reduction methods for brain-computer interface. *Information Technology and Control*, *48*(2), 225–234. https://doi.org/10.5755/j01.itc.48.2.23091

Kantarcioglu, M., & Shaon, F. (2019). *Securing Big Data in the age of AI* [Conference session]. Proceedings – 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA 2019, Los Angeles, CA, pp. 218–220. https://doi.org/10.1109/TPS-ISA48467.2019.00035

Khan, N., Naim, A., Hussain, M. R., Naveed, Q. N., Ahmad, N., & Qamar, S. (2019). *The 51 V's of Big Data: Survey, technologies, characteristics, opportunities, issues and challenges* [Conference session]. ACM International Conference Proceeding Series, Crete, Greece, Part F1481, pp. 19–24. https://doi.org/10.1145/3312614.3312623

Kim, J., & Park, N. (2020). A face image virtualization mechanism for privacy intrusion prevention in healthcare video surveillance systems. *Symmetry*, *12*(6), 891. https://doi.org/10.3390/SYM12060891

Koo, J., Kang, G., & Kim, Y. G. (2020). Security and privacy in Big Data life cycle: A survey and open challenges. *Sustainability (Switzerland)*, *12*(24), 1–32. https://doi.org/10.3390/su122410571

Krishna, R. R., Priyadarshini, A., Jha, A. V, Appasani, B., & Srinivasulu, A. (2021). State-of-the-art review on IoT threats and attacks : Taxonomy, challenges and solutions. *Sustainability*, *13*(16), 9463. https://doi.org/10.3390/su13169463

Kuhn, D. R., Walsh, T. J., & Fries, S. (2005). Security considerations for voice over IP systems recommendations of the national institute of standards and technology. *National Institute of Standards and Technology, 800–58*, 1–93.

Lee, W., Stolfo, S. J., Chan, P. K., Eskin, E., Fan, W., Miller, M., Hershkop, S., & Zhang, J. (2001). *Real time data mining-based intrusion detection* [Conference session]. Proceedings – DARPA Information Survivability Conference and Exposition II, Anaheim, CA, DISCEX 2001, *1*, pp. 89–100. https://doi.org/10.1109/DISCEX.2001.932195

Li, M., Zang, W., Bai, K., Yu, M., & Liu, P. (2013). *MyCloud – Supporting user-configured privacy protection in cloud computing* [Conference session]. ACM International Conference Proceeding Series, New Orleans, LA, United States, pp. 59–68. https://doi.org/10.1145/2523649.2523680

Liu, Z. C., Xiong, L., Peng, T., Peng, D. Y., & Liang, H. B. (2018). A realistic distributed conditional privacy-preserving authentication scheme for vehicular ad hoc networks. *IEEE Access*, *6*, 26307–26317. https://doi.org/10.1109/ACCESS.2018.2834224

Manjula, K., & Anandaraju, M. B. (2018). A comparative study on feature extraction and classification of mind waves for brain computer interface (BCI). *International Journal of Engineering and Technology(UAE)*, *7*(1), 132–136. https://doi.org/10.14419/ijet.v7i1.9.9749

McDermott, Y. (2017). Conceptualising the right to data protection in an era of Big Data. *Big Data and Society*, *4*(1), 1–7. https://doi.org/10.1177/2053951716686994

Mohan, K., Shrivastva, P., Rizvi, M. A., & Singh, S. (2014). *Big Data privacy based on differential privacy a hope for Big Data*. https://doi.org/10.1109/.167

Mujahid, A., Awan, M. J., Yasin, A., Mohammed, M. A., Damaševičius, R., Maskeliūnas, R., & Abdulkareem, K. H. (2021). Real-time hand gesture recognition based on deep learning YOLOv3 model. *Applied Sciences (Switzerland)*, *11*(9), 4164. https://doi.org/10.3390/app11094164

Nair, K. K., Helberg, A., & Van Der Merwe, J. (2016). *An approach to improve the match-on-card fingerprint authentication system security* [Conference session]. 2016 6th International Conference on Digital Information and Communication Technology and Its Applications, Konya, Turkey, DICTAP 2016, pp. 119–125. https://doi.org/10.1109/DICTAP.2016.7544012

Ninghui, L., Tiancheng, L., & Venkatasubramanian, S. (2007). *T-Closeness: Privacy beyond k-anonymity and ℓ-diversity* [Conference session]. Proceedings – International Conference on Data Engineering, 3, pp. 106–115, Istanbul, Turkey. https://doi.org/10.1109/ICDE.2007.367856

Onan, A. (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications*, *42*(20), 6844–6852. https://doi.org/10.1016/j.eswa.2015.05.006

Onan, A. (2019). Topic-enriched word embeddings for sarcasm identification. *Advances in Intelligent Systems and Computing*, *984*, 293–304. https://doi.org/10.1007/978-3-030-19807-7_29

Onan, A. (2021). Ensemble of classifiers and term weighting schemes for sentiment analysis in Turkish. *Scientific Research Communications*, *1*(1), 1–12. https://doi.org/10.52460/src.2021.004

Onan, A., & Korukoğlu, S. (2016). Exploring performance of instance selection methods in text sentiment classification. *Advances in Intelligent Systems and Computing*, *464*, 167–179. https://doi.org/10.1007/978-3-319-33625-1_16

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25–38. https://doi.org/10.1177/0165551515613226

Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, *53*(4), 814–833. https://doi.org/10.1016/J.IPM.2017.02.008

Onan, A., & Tocoglu, M. A. (2020). Satire identification in Turkish news articles based on ensemble of classifiers. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(2), 1086–1106. https://doi.org/10.3906/elk-1907-11

Onan, A., & Tocoglu, M. A. (2021). A term weighted neural language model and stacked bidirectional LSTM based framework

for sarcasm identification. *IEEE Access*, *9*, 7701–7722. https://doi.org/10.1109/ACCESS.2021.3049734

Patel, S. C., Graham, J. H., & Ralston, P. A. S. (2008). Quantitatively assessing the vulnerability of critical information systems: A new method for evaluating security enhancements. *International Journal of Information Management*, *28*(6), 483–491. https://doi.org/10.1016/j.ijinfomgt.2008.01.009

Peter, S. (2005). *Ripped by AaLl86. In security*.

Pham, V., & Dang, T. (2019, December). *CVExplorer: Multidimensional visualization for common vulnerabilities and exposures* [Conference session]. Proceedings – 2018 IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, pp. 1296–1301. https://doi.org/10.1109/BigData.2018.8622092

Pötzsch, S. (2009). Privacy awareness: A means to solve the privacy paradox? *IFIP Advances in Information and Communication Technology*, *298*(216483), 226–236. https://doi.org/10.1007/978-3-642-03315-5_17

Rajan, S., van Ginkel, W., & Sundaresan, N. (2012, November). Cloud security alliance (CSA): Top ten Big Data security and privacy challenges. *Csa*, *1*, 1–11.

Rastogi, N., Singh, S. K., & Singh, P. K. (2018, November 1). *Privacy and security issues in Big Data: Through Indian prospective* [Conference session]. Proceedings - 2018 3rd International Conference On Internet of Things: Smart Innovation and Usages, IoT-SIU 2018, Bhimtal, India. https://doi.org/10.1109/IoT-SIU.2018.8519858

Sánchez-Moreno, D., Batista, V. L., Vicente, M. D. M., Lázaro, Á. L. S., & Moreno-García, M. N. (2020). Exploiting the user social context to address neighborhood bias in collaborative filtering music recommender systems. *Information (Switzerland)*, *11*(9), 439. https://doi.org/10.3390/INFO11090439

Science, C. (2018). Comparison and evaluation of information retrieval models. *VFAST Transactions on Software Engineering*, *6*(1), 7–14. https://doi.org/10.21015/vtse.v13i1.496

Shamsi, J. A., & Khojaye, M. A. (2018). Understanding privacy violations in Big Data systems. *IT Professional*, *20*(3), 73–81. https://doi.org/10.1109/MITP.2018.032501750

Sharif, A., Soroya, S. H., Ahmad, S., & Mahmood, K. (2021). Antecedents of self-disclosure on social networking sites (SNSs): A study of facebook users. *Sustainability (Switzerland)*, *13*(3), 1–21. https://doi.org/10.3390/su13031220

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., Chen, S., Liu, D., & Li, J. (2020). Performance comparison and current challenges of using machine learning techniques in cybersecurity. *Energies*, *13*(10), 2509. https://doi.org/10.3390/en13102509

Shaukat Dar, K., & Ulya Azmeen, S. M. (2015). Dengue fever prediction: A data mining problem. *Journal of Data Mining in Genomics & Proteomics*, *6*(3), 1–5. https://doi.org/10.4172/2153-0602.1000181

Singh, M., Halgamuge, M. N., Ekici, G., & Jayasekara, C. S. (2018). A review on security and privacy challenges of Big Data. *Lecture Notes on Data Engineering and Communications Technologies*, *14*, 175–200. https://doi.org/10.1007/978-3-319-70688-7_8

Sivan, R., & Zukarnain, Z. A. (2021). Security and privacy in cloud-based e-health system. *Symmetry*, *13*(5), 742. https://doi.org/10.3390/sym13050742

Soria-Comas, J., & Domingo-Ferrert, J. (2013). *Differential privacy via t-closeness in data publishing* [Conference session]. 2013 11th Annual Conference on Privacy, Security and Trust, PST 2013, Tarragona, Spain, pp. 27–35. https://doi.org/10.1109/PST.2013.6596033

Soria-Comas, J., & Domingo-Ferrer, J. (2016). Big Data privacy: Challenges to privacy principles and models. *Data Science and Engineering*, *1*(1), 21–28. https://doi.org/10.1007/s41019-015-0001-x

Stouffer, K., Falco, J., & Scarfone, K. (2011). Guide to industrial control systems (ICS) security. *NIST Special Publication*, *800*(82), 16.

Stouffer, K., Falco, J., & Scarfone, K. (2011). GUIDE to industrial control systems (ICS) security. *The Stuxnet Computer Worm and Industrial Control System Security*, 11–158.

Sweeney, L. (2002). A model for protecting privacy. *Ieee Security and Privacy*, *10*(5), 1–14.

Ur Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. Y., & Khan, S. U. (2016). Big Data reduction methods: A survey. *Data Science and Engineering*, *1*(4), 265–284. https://doi.org/10.1007/s41019-016-0022-0

Varshney, S., Munjal, D., Bhattacharya, O., Saboo, S., & Aggarwal, N. (2020, December 16). *Big Data privacy breach prevention strategies* [Conference session]. Proceedings - 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security, ISSSC 2020, Gunupur Odisha, India. https://doi.org/10.1109/iSSSC50941.2020.9358878

Verizon. (2016). 2016 Data breach investigations report. *Verizon Business Journal*, *1*, 1–65.

Ward, R. M., Schmieder, R., Highnam, G., & Mittelman, D. (2013). Big Data challenges and opportunities in high-throughput sequencing. *Systems Biomedicine*, *1*(1), 29–34. https://doi.org/10.4161/sysb.24470

Wu, D., Yang, B., & Wang, R. (2016). Scalable privacy-preserving Big Data aggregation mechanism. *Digital Communications and Networks*, *2*(3), 122–129. https://doi.org/10.1016/j.dcan.2016.07.001

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Xiang, Y., Au, M. H., & Kutylowsky, M. (2016). Security and privacy in Big Data. *Concurrency Computation*, *28*(10), 2856–2857. https://doi.org/10.1002/cpe.3796

Xu, L., Wu, X., & Zhang, X. (2012). *Cl-Pre: A certificateless proxy re-encryption scheme for secure data sharing with public cloud* [Conference session]. ASIACCS '12: Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, Seoul, Korea, pp. 87. https://doi.org/10.1145/2414456.2414507

Yang, G., Luo, S., Zhu, H., Xin, Y., Xiao, K., Chen, Y., Li, M., & Wang, Y. (2019). A mechanism to improve effectiveness and privacy preservation for review publication in LBS. *IEEE Access*, *7*, 156659–156674. https://doi.org/10.1109/ACCESS.2019.2949452

Yasir, M., Afzal, S., Latif, K., Chaudhary, G. M., Malik, N. Y., Shahzad, F., & Song, O. Y. (2020). An efficient deep learning based model to predict interest rate using twitter sentiment. *Sustainability (Switzerland)*, *12*(4), 1660. https://doi.org/10.3390/su12041660

Yen, T. F., Oprea, A., Onarlioglu, K., Leetham, T., Robertson, W., Juels, A., & Kirda, E. (2013). *Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks* [Conference session]. ACM International Conference Proceeding Series, New Orleans, LA, United States, pp. 199–208. https://doi.org/10.1145/2523649.2523670

Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of Big Data. *IEEE Access*, *4*, 2751–2763. https://doi.org/10.1109/ACCESS.2016.2577036

Zhai, Y., Ong, Y. S., & Tsang, I. W. (2014). The emerging ? Big dimensionality? *IEEE Computational Intelligence Magazine*, *9*(3), 14–26. https://doi.org/10.1109/MCI.2014.2326099

Zhang, D. (2018). *Big Data Security and Privacy Protection* [Conference session]. Proceedings of the 8th International Conference on Management and Computer Science (ICMCS 2018), 77(Icmcs), Shenyang, China, pp. 275–278. https://doi.org/10.2991/icmcs-18.2018.56

Zhang, L., Stoffel, A., Behrisch, M., Mittelstädt, S., Schreck, T., Pompl, R., Weber, S., Last, H., & Keim, D. (2012). *Visual analytics for the Big Data era – A comparative review of state-of-the-art commercial systems* [Conference session]. IEEE Conference on Visual Analytics Science and Technology 2012, VAST 2012 – Proceedings, Seattle, WA, pp. 173–182. https://doi.org/10.1109/VAST.2012.6400554

Zhang, Q., Yang, L. T., & Chen, Z. (2016). Privacy preserving deep computation model on cloud for Big Data feature learning. *IEEE Transactions on Computers*, *65*(5), 1351–1362. https://doi.org/10.1109/TC.2015.2470255

Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, *28*(3), 583–592. https://doi.org/10.1016/j.future.2010.12.006

Zwitter, A. (2014). Big Data ethics. *Big Data and Society*, *1*(2), 1–6. https://doi.org/10.1177/2053951714559253