Journal of Computer Information Systems

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucis20

Systematic Survey: Secure and Privacy-Preserving Big Data Analytics in Cloud

Arun Amaithi Rajan & Vetriselvi V

To cite this article: Arun Amaithi Rajan & Vetriselvi V (2023): Systematic Survey: Secure and Privacy-Preserving Big Data Analytics in Cloud, Journal of Computer Information Systems, DOI: 10.1080/08874417.2023.2176946

To link to this article: https://doi.org/10.1080/08874417.2023.2176946

Published online: 24 Feb 2023.

JOURNAL OF COMPUTER INFORMATION SYSTEMS
https://doi.org/10.1080/08874417.2023.2176946

Systematic Survey: Secure and Privacy-Preserving Big Data Analytics in Cloud

Arun Amaithi Rajan and Vetriselvi V

Anna University, Chennai, India

ABSTRACT

Massive data are being generated exponentially every day. Therefore, analytics over data is inevitable nowadays to gain meaningful insights. Big Data Analytics (BDA) is powerful in critical applications while making effective decisions. Since the data to be processed are enormous in local systems, it is getting stored and processed in cloud platforms. Most of the clouds are public which are third-party resources. Security and privacy are the greatest concerns when it comes to the cloud. In the Big Data era, secure and private BDA has acquired the center of attention. This survey analyzes the various security and privacy solutions for BDA in the cloud environment from the combined perspectives of the three main subsystems: secure access control, secure data storage, and secure and private learning. Various techniques are studied and presented in each subsystem. The primary aim of this paper is to provide an overview of secure and private BDA in the cloud. Research challenges and future research directions in this area have also been presented.

KEYWORDS

Big data analytics; privacy and security; secure access control; secure data storage; secure and private learning; cloud computing

Introduction

We are living in an era of big data. 'Big data' is a general term for large and complex structured and unstructured data so that conventional data processing applications and systems cannot adequately handle them. Small-scale and homogeneous data are relatively easy to process, whereas processing large-scale heterogeneous data is tedious.

Tremendous data are being generated from various sources, such as the Internet of Things (IoT), Health care, Education, Banking, and Internet of Vehicles (IoV) domains. A single airplane could generate 10 TB of data in 30 min.[1] So, imagine the huge explosion of data. These data need not be grown linearly. Interestingly, it is booming in an exponential fashion. Big Data has the following characteristics (5Vs)[2] and is shown in Figure 1.

Volume – Describes how much data would be processed. Both structured and unstructured data are possible.

Velocity – Rate of data generation, which establishes the data's true potential.

Variety – Heterogeneous and diverse sources and the nature of the data.

Value – Usefulness of the gathered data.

Veracity – Describes degree of fineness of the data coming from various sources.

Raw data do not reveal any interesting hidden information or pattern in it. So, we need some data analysis. Processing huge data will help us gain useful insights from it, and the knowledge derived from the data helps us to make better decisions in critical applications.[3] We can call this process as Big Data Analytics (BDA). This involves data cleaning, data transformation, mining, and so on. Artificial intelligence (AI) learning algorithms play a major role in identifying complex patterns or deriving new insights from complex data. BDA helps in making quick and better decisions in every domain, and additionally, it contributes more to risk management with minimal cost. BDA is of different types, namely predictive, descriptive, prescriptive, and diagnostic analytics.[4]

Descriptive Analytics: This type of analytics provides an answer for what has occurred and manipulates raw data from many data sources to provide significant information about the past. These findings, however, merely point out as to what is correct or wrong without explaining the reason.

Diagnostic Analytics: This type of analytics is where historical data can be compared to other data to throw light on why something occurred and provide an in-depth understanding of a specific issue. At the same time, we need to have comprehensive data available; otherwise, data collection can be time-consuming and unique for each problem.

Predictive Analytics: This kind of analytics indicates what is most likely to occur (predict). It is a useful tool for forecasting since it uses the outcomes of descriptive

CONTACT Arun Amaithi Rajan arunamaithirajan@gmail.com Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

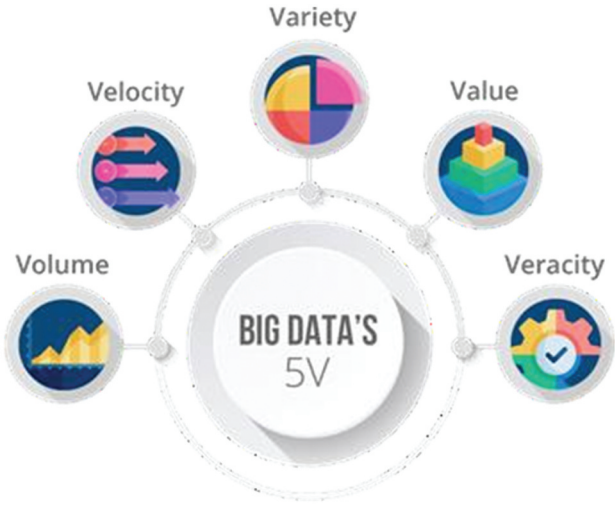

Figure 1. 5V's of big data.

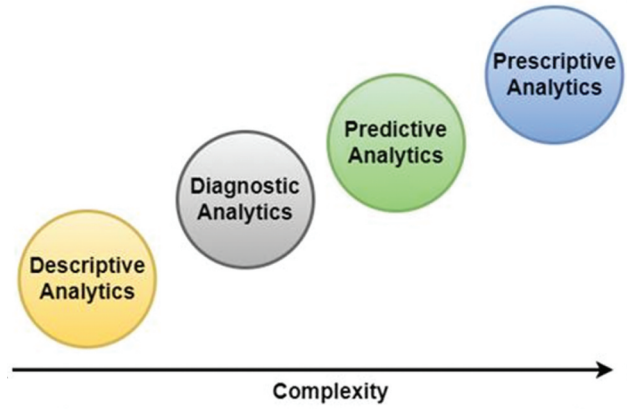

Figure 2. Types of BDA.

and diagnostic analytics to locate clusters and exceptions and to predict future trends. It falls under the category of advanced analytics. It offers several benefits, including sophisticated analysis based on deep learning (DL) and machine learning (ML), and the proactive approach that forecasts enable. However, as forecasting is merely an estimate, the accuracy of which is significantly controlled by the stability of the situation and the quality of the data, it involves careful preprocessing and regular optimization. Nowadays, predictive analytics is becoming the center of attention in the industry and in academia.

Prescriptive Analytics: This analytics combines cutting-edge tools and technology, such as ML, business rules, and algorithms, making it complex to deploy and administer. Its goal is to prescribe what action to take to solve a future problem or to fully capitalize on a promising trend. Figure 2 clearly shows the complexity level associated with each type of BDA.

When data grow larger, analytics using that data is not as easy as expected in local systems. It requires more memory space to accommodate big data and high processing systems to do analytics over big data. It requires large servers for which we rely on third parties. Here is where cloud computing comes into the picture. Cloud providers such as Google, IBM, and Microsoft provide more storage and distributed computational ability to handle the data efficiently.[5] Cloud services provide Software as a Service, Platform as a Service, Infrastructure as a Service, and so on that could be utilized by us to process high voluminous data simultaneously and efficiently.[6]

There are some issues in processing high voluminous data in the cloud, such as scalability, reliability, lack of data, and data and analytic security. In cloud systems, privacy and security have become the greatest concern.[7] More often, the big data getting analyzed is related to the person or to the things which should be protected in the specific context. Since trust is questionable in third-party cloud services, BDA in the cloud is leading to data leakage sometimes. To overcome this issue, we need secure and private BDA. Generally, confidentiality and privacy create some confusion between them. Tran et al.[8] clearly state that confidentiality is "data-oriented", whereas privacy is "data-owner-oriented."

Figure 3 depicts the general architecture of secure and private BDA in the cloud. In this architecture, there are three subsystems: access control (AC), data storage, and learning modules where security and privacy can be investigated. Secure AC (SAC) is an important subsystem that does a kind of entry-level restriction where attackers try to get into the system. Secure storage provides added security even if AC is breached and the data will be in an encrypted format in the cloud. Secure and private learning of BDA is the most crucial level where insights are learned without revealing the data owner's identity and confidentiality of the data. Thus, secure and private BDA in the cloud can be achieved by working on these three major subsystems. There are specific surveys on each subsystem.[3,9,10] However, it needs to be considered as a single architecture as shown in Figure 3 and developed together. This paper focuses on the state-of-the-art secure and private BDA in the cloud environment from the major focused area perspectives as mentioned in Figure 4.

In the big data era, privacy is more concerned with sensitive data. In search of recent research challenges and future directions in the prospective research area "security and privacy of BDA in the cloud," we are in need of a systematic and a comparative study on recent secure and private BDA techniques in the cloud. To the best of our knowledge, none of the surveys on secure and private BDA in the cloud have focused together on these three

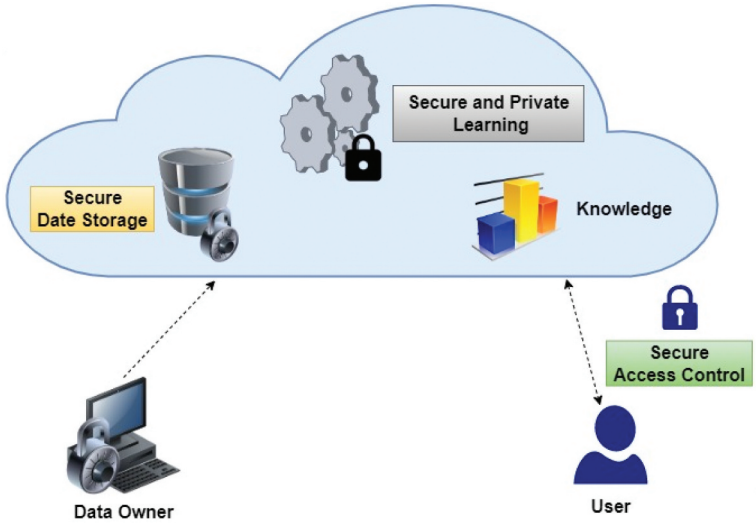

Figure 3. Secure and private BDA in the cloud.

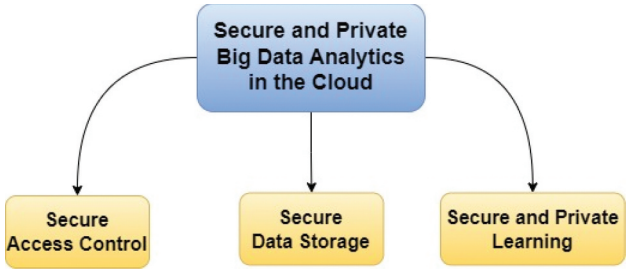

Figure 4. Focused research areas.

major areas: **SAC, secure data storage, and secure and private learning of BDA** in the cloud, which motivated us to survey secure and private BDA in the cloud.

The major contributions of this paper are as follows.

(1) Systematic survey has been conducted on secure and private BDA in the cloud.
(2) Recent research done on three major subsystems: SAC, secure data storage, and secure and private learning of BDA have been briefly summarized.
   a. The need for SAC with the most recent SAC mechanisms has been explained.
   b. The importance of secure data storage has been elaborated with different types of algorithms in it.
   c. Secure and private learning techniques involved in BDA are detailed.

(3) The scope for future research directions of secure and private BDA in the cloud has been outlined.

The rest of the paper is organized as follows: Section 2 explains the literature search methodology. Section 3 presents the existing techniques in SAC. Analysis of secure data storage has been detailed in section 4. Existing secure and private learning techniques have been elaborated in section 5. Section 6 discusses applications that require security and privacy. Future directions in this research domain are focused in section 7, and section 8 concludes the paper.

Our search methodology

We have conducted a brief study to complete this systematic survey. Finding research articles for this study is an iterative and time-consuming manual process. We adopted systematic literature review process for selecting and reviewing articles as explained in a research article by Snyder.[11] We fetched research and survey articles from popular publishers such as IEEE, Springer, arXiv, Elsevier, MDPI, and ACM, with the keywords such as "big data analytics," "security," "privacy," and "cloud" and during the period 2016–2022. We got around 300+ articles from our search. We have reviewed and picked 70+ papers manually as process shown in Figure 5.

There are two phases in the selection process: abstract screening and full article screening. While reviewing, we focused on the **scope, scientific depth, and novelty** of the article. After abstract screening, most of the articles are removed based on the inclusion criteria. Minimal amount of articles are removed after full article screening. Finally, the selected articles are categorized into three groups which help us to review and summarize them quickly. Table 1 shows the number of papers referred by each publisher.

The pie chart in Figure 6 shows the recentness of articles and our survey. For this chart, we excluded the literature which has been used to cite the definition of

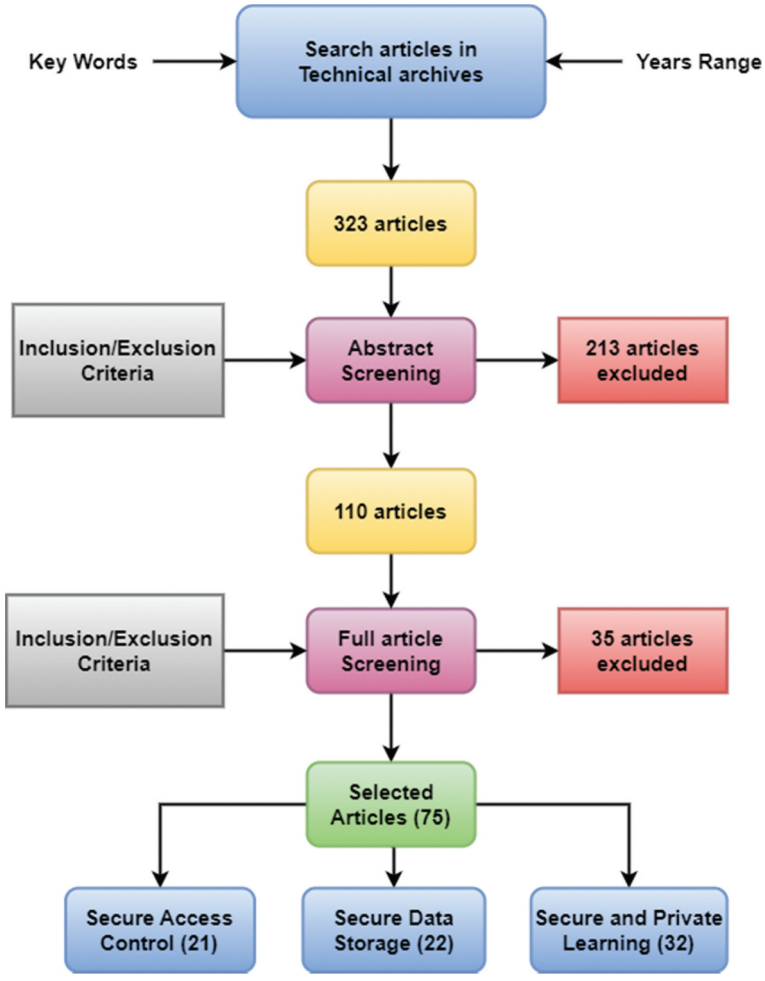

Figure 5. Research articles selection process.

Table 1. Publishers and papers count.

Publisher	Papers count
IEEE	23
Springer	17
Elsevier	10
MDPI	10
arXiv	9
ACM	7
Others	16

some technical terms. There exists a simple study on privacy and security in the general big data paradigm[12] which has taken literature until 2016. This prompted us to choose articles from 2017. Specifically, around 70% of the articles are in the period 2020 to 2022 for our survey. In each subsystem: SAC, secure data storage, and secure and private learning of BDA, we have taken survey articles and recent research articles.

Figure 7 explains the categories of recent techniques in each subsystem. This shows the summary of the conducted survey.

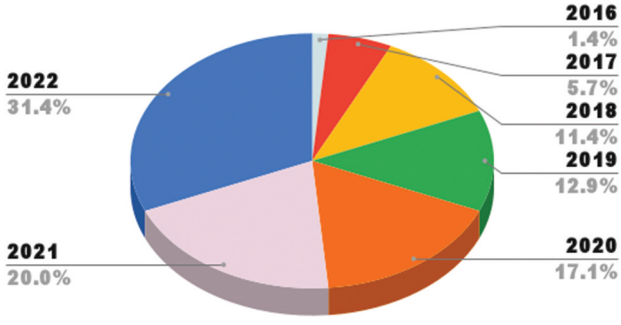

Figure 6. Recentness of the survey.

Secure access control

This section briefly analyzes the recent SAC mechanisms in the cloud. AC is a basic protection for applications in the cloud. Bypassing SAC is a major attack technique used by attackers. So, secureness of AC should be kept with high priority. AC in cloud security

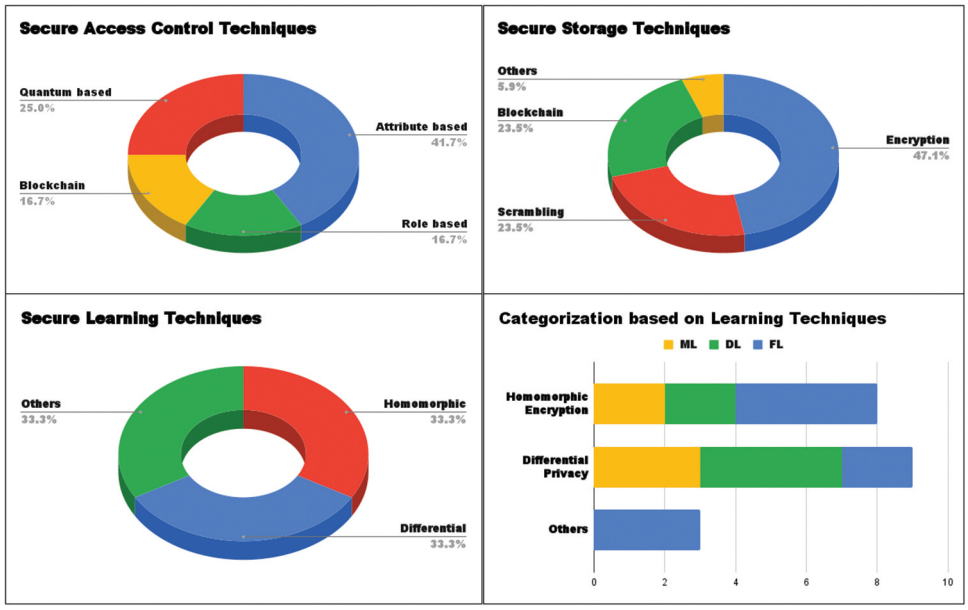

Figure 7. Articles categorization summary.

is a system with which an organization can regulate and monitor permissions, or access to their data by formulating various policies suited chosen by the organization.[9] AC in cloud security helps organizations gain macro-level visibility into their data and user behavior.

Reduced security risk from unauthorized access to physical and logical systems is the main aim of AC. AC guarantees that security technology and AC policies are in place to secure sensitive data, such as customer data, and is a critical part of security compliance programs. Access to networks, computer systems, applications, files, and sensitive data, including Personally Identifiable Information and Intellectual Property, is typically restricted by architecture and policies within organizations.

In dynamic IT environments involving both cloud services and on-premises systems, AC solutions can be complicated to manage. Technology providers have switched from single sign-on solutions to unified access management, which provides access restrictions for on-premises and cloud settings, as a result of some high-profile breaches.[13] Figure 8 shows a clear overview of SAC in cloud applications.

There are more research studies[14–16] actively going on in the area of SAC. El Sibai et al.[9] critically reviewed the SAC mechanisms for the cloud and categorized as Mandatory AC, Identity-based AC, Attribute-based AC (ABAC), and Role-based AC (RBAC). The authors explained the mechanisms in detail with scenarios that

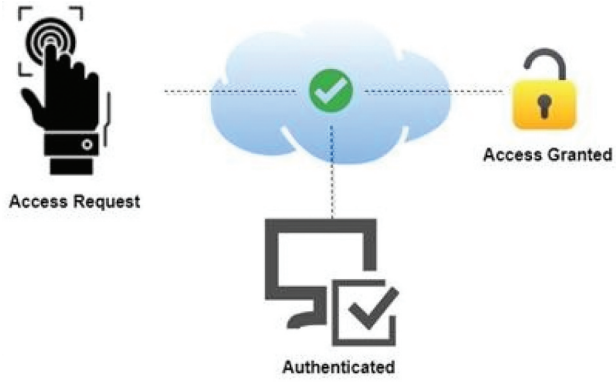

Figure 8. Overview of secure access control in the cloud.

made it easier to grasp. Comparison of AC mechanisms is evaluated using the following criteria: separation of duties, least privilege principle, policy management, scalability, capability delegation, safety, and flexibility of configuration. In another article, Sahi et al.[17] interestingly surveyed the state-of-the-art existing secure and privacy–preserving AC mechanisms in e-healthcare and presented ABAC and RBAC in health-care environment with different scenarios.

Based on the analysis of the latest articles on SAC techniques, four major techniques are being utilized in recent days as depicted in the following Figure 9. These techniques are briefly explained in the following subsections. Blockchain (BC) technology has occasionally been integrated with other AC techniques nowadays.

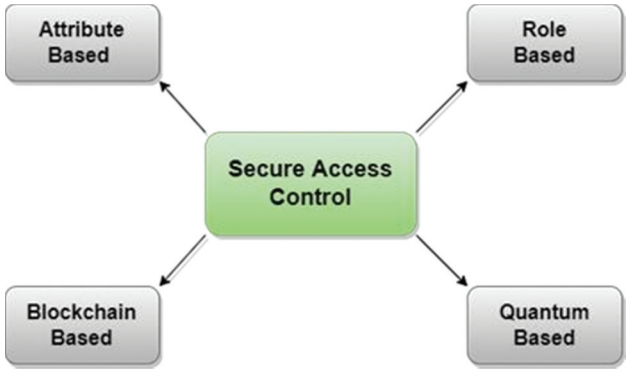

Figure 9. Secure access control techniques.

Attribute-based access control

Based on the user's or the resource's attributes, the ABAC model approves or rejects a user's request to access confidential systems.[18] More attributes are included in access structures, including authentication level, user qualifications, location, and time. On the internet, there is a vast amount of private information kept by outside parties where security is in doubt. Storing the data in an encrypted form is one way to address this issue and guard the data against any unauthorized usage. This solution's biggest drawback is that it forbids sharing encrypted data at a finer level.

To solve the above problem, attribute-based encryption (ABE) is being introduced. ABE is a kind of public-key encryption where the ciphertext and the user's secret key are reliant on attributes. In such a system, a ciphertext can only be decrypted if the user key's set of attributes coincides with those of the ciphertext. Goyal et al.[19] introduced ABE, specifically Key-Policy ABE (KP-ABE) for fine-grained AC of encrypted data in 2006 and Waters et al. introduced Ciphertext-Policy ABE (CP-ABE) in 2007.[20] Since ABE is advantageous in the IoT environment, it helps to store data securely in untrusted storage. Rasori et al.[21] deeply investigated 60+ ABE schemes that are suitable for the IoT environment. Comparative analysis has been done on access structure's expressiveness and on key performance indicators, such as producer bandwidth overhead, key authority bandwidth overhead, and CPU load. The authors projected drawbacks such as traceability, accountability, puncturability, and post-quantum security of ABE in IoT.

Figure 10 explains KP-ABE and CP-ABE in the health-care scenario. In KP-ABE, the data owner creates a ciphertext that is encrypted with the user's attribute who is capable of accessing the data and sent to the cloud storage. After that, a user who needs the data can get the ciphertext by accessing the cloud. The user can then decrypt the ciphertext by producing a key to decrypt the ciphertext and building an access structure depending on their attributes. For instance, the data owner encrypts the data and uploads it to the cloud with the attributes [Nurse, Doctor]. The user with the [Doctor] or [Nurse] attributes can create an access structure so that the user can generate a key to decrypt the ciphertext.

In both CP-ABE and KP-ABE, the data owner can specify who has access to the data. The access structure and key are generated differently depending on who does so. In contrast to KP-ABE, which relies on the user to generate a key directly, CP-ABE needs an authority that can issue keys after confirming the user attributes. CP-ABE attributes are used to define a user's credential, and the encryptor decides who can decrypt the data. KP-ABE attributes are used to describe the encrypted data, and policies are embedded into the user's keys. Recent techniques are incorporated with the traditional ABE schemes are explained below.

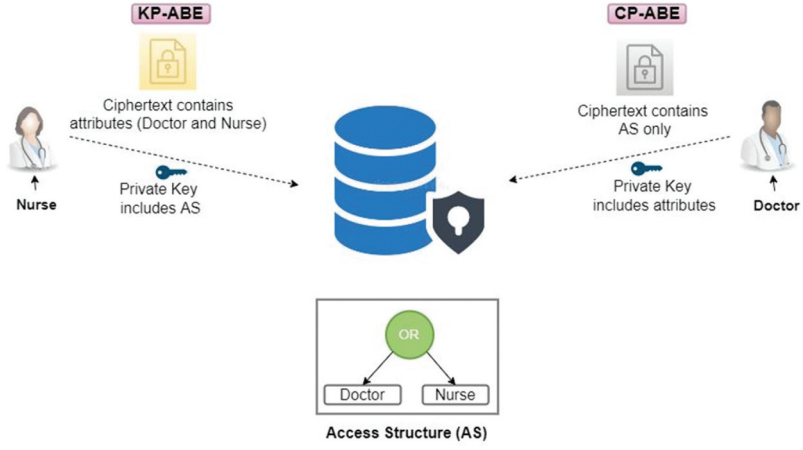

Figure 10. KP-ABE and CP-ABE.

Traditional CP-ABE with enhanced security was implemented by Chinnasamy et al.[14] They make use of SHA1 to secure the policy inside the ciphertext. Hiding access policy using SHA1 within the ciphertext is the greatest contribution here in this article. Ciphertext size also gets reduced since a hash is computed. Computational costs of existing CP-ABE schemes are analyzed with the proposed CP-ABE scheme and proved its efficiency.[14] Hwang et al.[22] focused on the key abuse issue in the CP-ABE scheme where the user can share their keys with others which let them access the cloud server with legitimate keys. The authors proposed a solution by verifying the user identity with a tracking facility. Verifiable outsourcing is required for this scheme to achieve SAC and user anonymity in key distribution. System has been simulated using the medical data sharing in the cloud with Internet of Medical Things (IoMT).

Wang et al.[23] developed a secure cloud storage framework with AC based on CP-ABE over Ethereum BC technology. This framework provides a decentralized AC mechanism to avoid major threats. Multi-authority ABE is exploited, and Ethereum is utilized to convert the traditional centralized ABAC to distributed ABAC. The performance of the designed framework is expressed in terms of execution time plotted against a number of attributes. The developed framework[23] runs 4× times more than the general CP-ABE. This system brings development in both security and performance of CP-ABE.

New AC with privacy preservation using CP-ABE and local differential drivacy (LDP) was introduced by Song et al.[24] This takes ABE to the next level (Differential Privacy (DP) is explained in the later section). They enhanced the privacy level and the performance using perturbation and encryption with CP-ABE. A comparison of synthetic datasets and real-time datasets has been done to evaluate the performance of privacy AC. The privacy rate of AC is improved by adding LDP.

Yang et al.[25] proposed a self-adaptive AC in health care that handles normal and emergency medical situations efficiently. They utilized the ABE schemes to provide AC and introduced a break-glass access mechanism that handles emergency medical conditions and cross-domain data sharing. The goal of a break-glass access methodology is to enable the recovery of all patient history medical records using a password-based break-glass key. It is evident from the analysis that ABE schemes are adaptable and can be used in conjunction with cutting-edge technologies to offer AC. As CP-ABE is more informative and efficient, it is more utilized. Issues in ciphertext, key, security, and performance of CP-ABE are focused, and solutions are proposed using BC and LDP in recent years.

Role-based access control

The main idea behind RBAC is to provide restricted access to the system depending on the role of a user within an organization. RBAC is one of the principal methods in advanced AC.[26] The levels of access that users have to the application are described by the roles in RBAC. Access can depend on many things, like authority, accountability, and job proficiency. Access to computer resources can also be restricted to particular operations, such as the capability to view, create, or alter a file.

Because of this, lower-level users typically do not have access to sensitive information if they do not require it to carry out the tasks. RBAC will assist in protecting sensitive data and vital applications in the organization. Figure 11 resembles a situation in which RBAC is used. Student-grade database is a sensitive storage which should be only modifiable by teaching and admin faculties. Students should be only able to view the grades. This scenario at education institutes is achieved via RBAC.

There exist more techniques to provide access based on the roles. Alshammari et al.[27] combined two AC mechanisms (task-based and role-based) and came up with new cryptographic task-role-based AC. A secure and private cloud system is designed based on trust between the owner and the consumer. Based on the trust, a particular "Task T" can be executed by a particular user with "Role R" in this system. The trust model is verified with multiple attacks such as Sybil and collude. Kim et al.[28] argued that RBAC added with BC is more efficient in video surveillance. Access is restricted via the role of the user, and integrity is maintained in a BC network. Simple RBAC mechanisms are being incorporated with the latest technologies to enhance security.

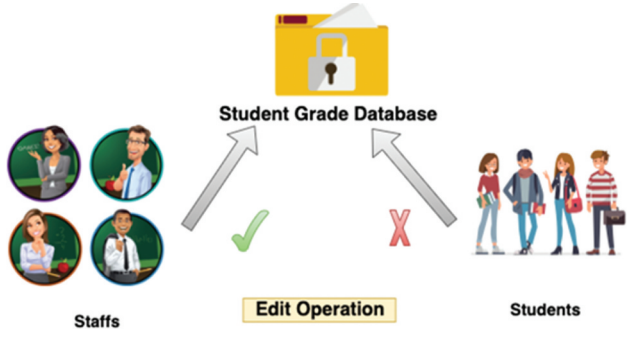

Figure 11. Role-based access control.

Blockchain-based access control

This section of the article concentrated on AC, which is majorly based on BC. Generally speaking, BC is a mechanism for storing data in a way that makes system modifications, hacking, and cheating difficult or impossible. Nakamoto[29] established the BC model in 2009. A BC is a digital database of transactions that is duplicated and dispersed among a network of computers. Whenever a new transaction starts on the BC, a copy of that transaction is inserted into each and every participant's ledger. Each block on the chain comprises some transactions. The term Distributed Ledger Technology (DLT) describes a decentralized database maintained by several people.

BC is another type of DLT in which transactions are stored with an irrevocable cryptographic signature known as a hash. To breach a BC system, hackers would need to change every block in the chain across all distributed versions of the chain. BCs like those used by Bitcoin and Ethereum are constantly expanding as new blocks are added to the chain, greatly enhancing the security of the ledger.[30] Figure 12 displays the BC's fundamental structure.

Recently, Deep et al.[31] offered a robust and interesting authentication method using BC . So, it is more difficult for an insider process (who is working in the organization) to alter user login information throughout user authentication. Due to the distributed ledger-based authentication mechanism, the insider is unable to access the user authentication data. Using unique IDs and signatures, all the users (insider and outsider) are securely verified. Insider activity can be tracked, and the activity logs cannot be altered as they are stored in the BC . Additionally, the cloud database's user AC is authenticated. Their mechanism[31] is tested using the Scyther formal system tool for resistance to denial-of-service and no-replay attacks.

In health care a controllable BC-based Electronic Health Record (EHR) sharing scheme has been proposed by Li et al.[15] BC is combined with interplanetary file systems (IPFS) to provide efficient and secure EHR sharing. IPFS-based EHR management is implemented for decentralized sharing of all EHRs and BC-based EHR management for secure abstract accessing. The public key of the patient is used to encrypt the EHR abstracts. AC over the EHR is in the data owner's hand. The BC is used as an intermediate authentication system to reroute the requester to the actual EHR. As BC is tamper proof, it provides data integrity while SAC. The next subsection focuses on how quantum-based techniques are used for SAC.

Quantum-based access control

The majority of authentication mechanisms in classical cryptosystems are based on cryptographic primitives. For instance, RSA and ElGamal are cryptosystems built on factorization or discrete logarithm hard problems. It is widely assumed that such primitives are vulnerable to quantum algorithms. Shor's algorithm[32] is a quantum algorithm that solves discrete logarithm and factorization problems in sub-exponential time complexity. So, if we continue to have the same cryptosystem, there is a chance that quantum computers will attack secure applications within polynomial time in the future. So, with a longer vision, more works are being done in quantum cryptography.[33] As we have hard problems in traditional systems, such as discrete logarithm, there are some hard problems in the quantum environment also such as lattice problems which prompted scientists to develop quantum cryptography algorithms. A survey on lattice-based cryptography implementations in software and hardware by Nejatollahi et al.[34] gives an overview of existing algorithms in post-quantum cryptography based on lattice problems. The authors worked on implementations which tackles different issues such as memory footprint, energy, security and, given some proposals for lattices in information security

Qiu et al.[35] utilized the quantum cryptography concepts to perform authentication in cloud-based applications. They showed the three-phase protocol: key

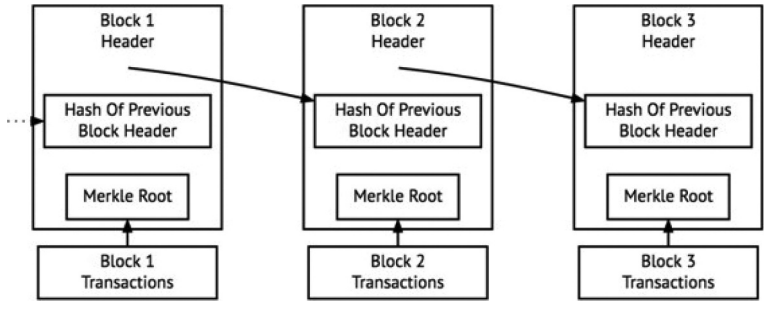

Figure 12. Basic structure of blockchain.

distribution, identity authentication, and digital certification. Subsequently, Huang et al.[36] proposed a group-based authentication protocol based on the lattice. Generally, lattice-based authentication uses N^{th} degree Truncated polynomial Ring Units (NTRU) asymmetric algorithms which are expensive. However, symmetry is used for group authentication in the proposed protocol. It has reduced time complexity and is secure against attacks, such as man-in-the-middle attacks, and replay attacks.[36]

Recently, Akleyak et al.[16] introduced a new lattice-based authentication mechanism for the IoT environment. The authors proposed a scheme that depends on the Inhomogeneous Short Integer Solution (ISIS) problem in lattices. Theoretical security attack analysis and performance comparison using computation time analysis have been done. Fu et al.[33] did a focused survey on lattice-based expressive ABE. Since bilinear-based ABE will not be effective in the quantum era, lattice-based ABE is expected to be in the future, and already research is going on in this area. Multiple schemes expressed from lattices have been discussed and compared. This gives us a lot of new directions to utilize ABE in lattice spaces.

The latest SAC techniques are discussed so far. In ABAC, CP-ABE is utilized for its advantages and issues, such as key abuse and ciphertext security, are focused. RBAC technique has been enhanced with BC. Solutions using LDP and BC are provided for SAC. With greater vision, quantum-based AC also involved in SAC. Traditional SAC techniques are trying to get converted using quantum technologies. A summary of the discussed recent AC techniques is shown in Table 2. Generally, the performance of AC mechanisms is evaluated using time complexity. The secureness of the mechanisms is evaluated by performing the attacks over the proposed schemes and proving robustness.

Legend of Table 2 is as follows: **AB**: Attribute-Based Access Control, **RB**: Role-Based Access Control, **BB**: Blockchain-based Access Control, **QB**: Quantum-Based Access Control, **Enc**: Whether encryption is involved or not? **SA**: Whether Security Analysis is done or not? **PA**: Whether Performance Analysis is done or not?

Secure data storage

This section is dedicated to secure data storage in cloud systems. Generally, big data will be stored in the cloud before analytics. First, types of big data are classified and then securing techniques are detailed.

Big data types and their security

We produce an unbelievable amount of data every second as the internet age develops. By 2025, the amount of data on the internet are anticipated to reach 163 zettabytes.[37] There are numerous tweets, selfies, purchases, e-mails, blog entries, and other types of digital data. This big data can be divided into three categories: structured, unstructured, and semi-structured.[38] It is shown in Figure 13.

Structured data

Structured data are easy to examine and sort since they have predetermined organizational characteristics and are present in the structured or tabular schema. Additionally, each field is distinct due to its predetermined nature and can be accessed alone or jointly with information from other fields. Because of that, structured data seem to be very beneficial and they enable quick data collection from diverse database locations.

Unstructured data

Unstructured data refer to information that is without any predetermined conceptual definitions and is difficult for traditional databases or data models to interpret or analyze. The majority of big data are made up of unstructured data, which include facts, dates, and numbers. Examples of this kind of big data include satellite imaging, mobile activities, audio, and video files. The amount of unstructured data are expanding as a result of the Instagram photos and YouTube videos we publish and watch.

Table 2. Summary of the recent SAC techniques.

Year	Article	AB	RB	BB	QB	Enc	SA	PA
2019	Yang et al.[25]	**Yes**	No	No	No	**Yes**	Yes	Yes
2019	Wang et al.[23]	**Yes**	No	**Yes**	No	No	Yes	Yes
2020	Hwang et al.[22]	**Yes**	No	No	No	**Yes**	Yes	Yes
2022	Chinnasamy et al.[14]	**Yes**	No	No	No	**Yes**	Yes	Yes
2022	Song et al.[24]	**Yes**	No	No	No	**Yes**	Yes	Yes
2021	Alshammari et al.[27]	No	**Yes**	No	No	No	Yes	Yes
2022	Kim et al.[28]	No	**Yes**	**Yes**	No	No	Yes	Yes
2019	Deep et al.[31]	No	No	**Yes**	No	No	Yes	Yes
2022	Li et al.[15]	No	No	**Yes**	No	No	Yes	Yes
2018	Qiu et al.[35]	No	No	No	**Yes**	**Yes**	Yes	Yes
2018	Huang et al.[36]	No	No	No	**Yes**	**Yes**	Yes	Yes
2022	Akleylek et al.[16]	No	No	No	**Yes**	**Yes**	Yes	Yes

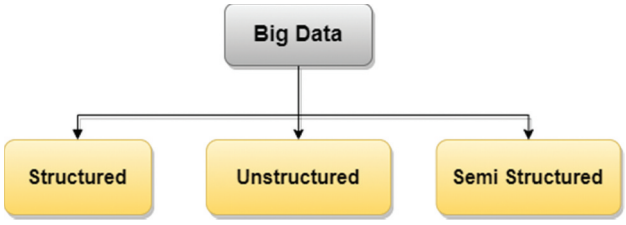

Figure 13. Types of big data.

Semi-structured data

A combination of structured and unstructured data is semi-structured data, which means that it inherits a few characteristics of structured data and unstructured data. For instance, JSON and XML are semi-structured data.

The big data which we have collected has to be securely stored in the cloud since we rely on a third party for storage and analytics. Multiple mechanisms[39–41] are available to protect the data in the cloud. The survey article presented by Cunha et -al.[38] showed the data type classification and privacy-preserving techniques for heterogeneous data such as encryption and obfuscation in detail. They proposed a privacy taxonomy that establishes a relation between data types and privacy-preserving mechanisms with privacy tools. The latest survey on general privacy-preserving techniques for data publishing in the cloud by Carvalho et al.[42] focused on perturbative, non-perturbative, and de-associative techniques. The authors discussed technical solutions to the privacy issues in big data and did a tool exploration. Open issues such as dynamic data, big data, and distributed data privacy are projected at last.

In our paper, we are mostly focusing on the secure storage techniques of unstructured data. Because of the tremendous proliferation of digital images, organizations and people are increasingly turning to the cloud for image storage and computing. However, unencrypted upload increases the potential of privacy leakage, whereas basic encryption impedes the efficient use of data. Particularly sensitive images need privacy and security. Image data's secure storage techniques are surveyed by Kaur et al.[43] extensively with comparative analysis. They had taken 14 general image encryption evaluation metrics and compared those metrics with each image encryption technique such as chaotic map, elliptic curve, DNA coding, and fuzzy logic. After analyzing the pool of the literature on secure storage techniques in the cloud, we have split the unstructured data's securing techniques into three categories as shown in Figure 14.

Each technique has been explained in the following subsections with the work done in recent years.

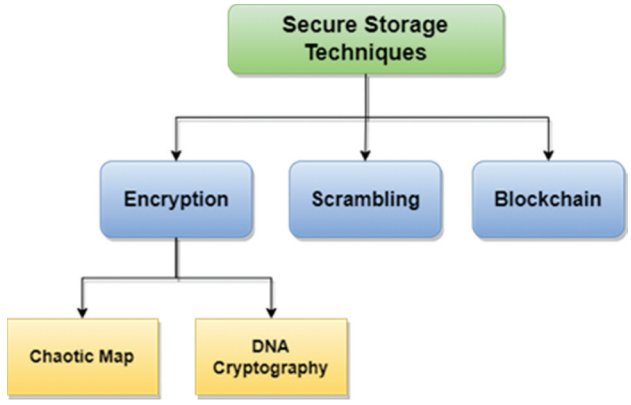

Figure 14. Secure storage techniques.

Encryption

Encryption is a technique where the original data are converted into encrypted data which can be only retrieved back with the proper key. There are many image encryption techniques available as discussed in Kaur et al.[43] We discuss spatial domain encryption techniques: chaotic maps and DNA cryptography in the following subsections.

Chaotic maps

The chaotic map is a field of study in mathematics that deals with dynamic systems exhibiting chaotic behavior. Every trivial change in the initial conditions of the system can produce substantial changes in the output. These maps can be either discrete or continuous.[44] As these chaotic maps show randomness, they perform well in pixel confusion of image encryption. For example, Figure 15 shows the bifurcation diagram of logistic map. When r ϵ (3.5,4], map becomes chaotic.

In 2016, a secure and fast image encryption algorithm had been devised with S-boxes and hyperchaotic mapping by Abduljabbar et al.[45] Their mechanism goes as follows: S-boxes generation, scrambling, and encryption. Different attack simulations had been simulated against the proposed mechanism. Instead of a single logistic map, a double logistic map had been used for image encryption by Pan et al.[46] and produced better accuracy. In 2020, Wan Y et al.[47] combined a hyperchaotic Qi system and a one-dimensional chaotic map to come up with a new image encryption scheme.

Huang et al.[48] employed a 3D non-equilateral Arnold transformation and hyperchaotic system to devise a novel bit-level encryption algorithm for images. A perturbing encryption method with homomorphism is detailed in Cheng et al.[39] In the encrypted domain, homomorphism assures information security and the

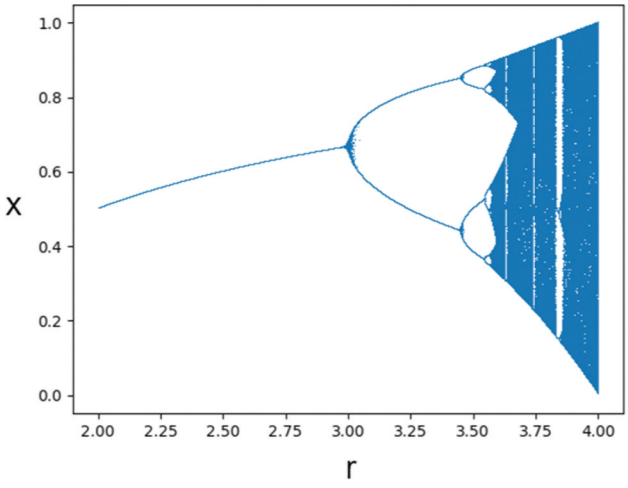

Figure 15. Logistic map – bifurcation diagram.

preciseness of discrete wavelet transformation. This approach allows the watermarking procedure to be executed in a risky outsourced environment while producing a watermarking effect comparable to the plaintext equivalent.

Since health care become increasingly important and concentrated, Ahmed et al.[49] specifically focused on medical images and applied "encryption and then compression" technique to achieve better accuracy. In an image, the proposed method preserves information which is crucial and necessary for correct diagnosis and significantly reduces the image size. Devised method[49] provides security to enable telemedicine application of e-Health services.

DNA cryptography

Data hiding methods using the deoxyribonucleic acid (DNA) sequence are known as DNA cryptography. In this cryptographic process, each letter of the alphabet is transformed into a unique combination of the four bases that make up human DNA: adenine (A), guanine (G), cytosine (C), and thymine (T).[50] Table 3 shows the eight DNA encoding rules.

Moosavi et al.[40] proposed an efficient and secure mutual authentication scheme based on the Elliptic Curve Cryptography (ECC) and the Quark lightweight hash design with added DNA Cryptography. Bao et al.[51] also designed an image encryption algorithm combining compressed sensing and DNA coding, which achieves better performance and more security. Both research studies used DNA encoding for pixel substitution.

Scrambling

This section presents how scrambling technique is used in secure image storage in the cloud. Data scrambling is a technique of altering or obfuscating sensitive information, a technique typically used to protect data's secrecy. In an image, usually, scrambling is done by shifting pixels, permutations, etc.

A safe image encryption system with a scrambling method based on Compressed Sensing (CS) is developed by Choi et al.[52] They adopted a sparse measurement matrix for efficient encryption, with non-zero elements generated by a key stream generator based on an Linear Feedback Shift Register (LFSR). Then, for diffusion, additional pairs of data scramblers, likewise based on LFSR, are attached behind the CS encryption. A new word-oriented feedback shift register technique was utilized in Amit Arora et al.[53] to achieve confidentiality of data with better performance.

The robust image steganography is focused on the work done by Sukumar et al.[54] which exploits Laplacian pyramid, Arnold scrambling, Redundant Integer Wavelet Transform (RIWT), and histogram shifting algorithm to facilitate secure communication of secret images in the medical context. Binary map, block, and pixel permutation with order-preserving encryption is being utilized in Xia et al.[55] which preserves the privacy and local binary pattern features of the images. These scrambling techniques are tested against chosen-ciphertext attacks and histogram analysis and proved robustness.

Blockchain and other techniques

In addition to encryption and scrambling technologies, a variety of solutions are available for secure cloud data storage. This subsection explains a few of these techniques. In most of the secure publishing schemes using BC, it has been used as a place where data can be kept securely as in a database. Ali et al.[56] introduced a trustworthy audit log management system using BC for cloud computing to prevent any malicious activities even by administrators. Their proposed system satisfies

Table 3. DNA encoding rules.

Bits	I	II	III	IV	V	VI	VII	VIII
00	A	A	G	C	G	C	T	T
01	G	C	A	A	T	T	G	C
10	C	G	T	T	A	A	C	G
11	T	T	C	G	C	G	A	A

Table 4. Brief summary of secure data storage techniques.

Year	Article	Enc	Scr	BC	Technique	SA	PA
2018	Pan et al.[46]	**Yes**	No	No	Double Logistic Map	Yes	Yes
2020	Huang et al.[48]	**Yes**	No	No	Hyperchaotic Map	Yes	Yes
2020	Wan Y et al.[47]	**Yes**	No	No	Hyperhaotic Qi Map	Yes	Yes
2022	Abduljabbar et al.[45]	**Yes**	No	No	Hyperchaotic Map	Yes	Yes
2022	Cheng et al.[39]	**Yes**	No	No	Watermark	Yes	Yes
2022	Ahmad et al.[49]	**Yes**	No	No	Compression	Yes	Yes
2022	Moosavi et al.[40]	**Yes**	No	No	ECC and DNA	Yes	Yes
2022	Bao et al.[51]	**Yes**	No	No	CS and DNA	Yes	Yes
2021	Sukumar et al.[54]	No	**Yes**	No	Arnold Scrambling	Yes	Yes
2021	Xia et al.[55]	No	**Yes**	No	Permutations	Yes	Yes
2022	Choi et al.[52]	No	**Yes**	No	LSFR	Yes	Yes
2022	Arora et al.[53]	No	**Yes**	No	WSFR	Yes	Yes
2020	Vetriselvi et al.[57]	No	No	**Yes**	CNN	Yes	Yes
2021	Ali et al.[56]	No	No	**Yes**	-	Yes	Yes
2021	Alquaralleh et al.[58]	No	No	**Yes**	ECC and DBN	Yes	Yes
2022	Alhazmi et al.[59]	No	No	**Yes**	Fragmentation	Yes	Yes
2021	Liu et al.[60]	No	No	No	Differential Privacy	Yes	Yes

correctness, forward security, and protection against corruption.

Vetriselvi et al.[57] devised a simple mechanism to store sensitive medical information in the BC . The heartbeat dataset is taken and classified as critical and noncritical beats, and which patient needs surgery has been inferred from that classification. Those patient's detailed EHR is maintained in the BC which prevents tampering of data and provides forgery prevention. Alqaralleh et al.[58] also secured IoMT data with the help of BC technology. Instead of storing the sensitive data in an encrypted form, a secure hash of those sensitive data has been stored in BC-enabled cloud storage. ECC is utilized for encryption, and the Deep Belief Network (DBN) is exploited to classify and predict the disease based on IoMT data. The performance of the proposed technique is measured by the retrieval speed of the encrypted data in a BC. A secure framework for big data by using fragmentation and BC technology is implemented by Alhamzi et al.[59] First, Big Data is classified first as nonsensitive, low sensitivity, and highly sensitive data; then if the data are highly sensitive, it gets fragmented, encrypted, and stored in the BC.

It is known that DP (explained in the later section) is an undisputed criterion that guarantees provable privacy. Due to the lack of a clear qualification on the meaningful difference between any two images, the application of DP on unstructured data such as images is not trivial. The article by Lio et al.,[60] for the first time, introduced a DP image which is a novel notion of image-aware DP. DP image can protect user's personal information in images, from both human and AI adversaries just by adding noise in the image feature vector.

In secure storage, there are two types. One category where secure operations are done over raw data and secured in insecure storage as encryption and scrambling techniques. Another type of raw data is getting stored in secure storage platforms such as BC. Both methods are secure where secure storage will be more expensive and retrieval operations are tedious. So, most of the cloud platforms utilize the first method. Table 4 briefly summarizes the secure storage techniques discussed so far in this section. To analyze the secureness of the storage mechanism, multiple attack analyses are available such as noise attack analysis, differential attack analysis, and so on. The performance of the mechanism is evaluated using histogram analysis and speed analysis.

Legend of Table 4 is as follows: **Enc**: Encryption, **Scr**: Scrambling, **BC**: Blockchain, **Technique**: Exact technique used within the category, **SA**: Whether Security Analysis is done or not? **PA**: Whether Performance Analysis is done or not?

Secure and private learning techniques

Big Data at rest will not provide us with trends. It needs to be analyzed and visualized to extract interesting patterns from it. There are four different kinds of analytics in Big Data as explained in the introduction: predictive, descriptive, prescriptive, and diagnostic.[4] Predictive analytics is our research area in focus and is becoming a trending analytics technique. Typically, data analytics involves Data Mining (DM) and AI techniques when it comes to Big Data. In AI, we have ML and DL techniques for learning patterns and for finding hidden information.[3] This section is dedicated to secure and private learning techniques of BDA in the cloud.

We are using millions of data for learning, but are we protecting those while training? **"No"**. The security of the model is questionable when it comes to the learning of sensitive data in critical cloud applications, such as medical imaging, smart city, IoV, etc. Figure 16 shows

the learning techniques that we are interested in and which are widely used in predictive analytics in the cloud.

In our survey, we focused on securing learning techniques ML, DL, and federated learning (FL) that are mostly used for BDA. Some surveys talk about the security and privacy of BDA in the cloud. Tran et al.[8] comprehensively conducted a scenario-based survey on private data analytics and proposed a taxonomy of privacy-preserving BDA in the cloud along with open future research directions. Pramanik et al.[10] also critically analyzed secure and privacy-preserving BDA methodologies. They also developed a new four-dimensional framework for studying and building the next generation of privacy-preserving BDA solutions. Recently, Yin et al.[61] surveyed privacy-preserving federated learning (PPFL) and designed a 5W scenario-based taxonomy that helps to categorize PPFL techniques.

After referring to multiple research articles on secure and private BDA,[62–64] we survey how these learning techniques can be secured and the privacy of the model can be preserved. Our studies focused on two major techniques: homomorphic encryption (HE) and DP in secure and private learning. A brief overview of the learning techniques is given before discussing how secure and privacy techniques are applied to learning modules.

The term AI refers to a collection of tools and technologies that can analyze data, categorize it, and make intelligent decisions based on specified factors. AI's main goal is to make robots learn and think like people while also improving their problem-solving abilities. Simply expressed, ML and DL are subsets of AI, whereas AI is a general term.

Learning techniques

Machine learning

The main idea of ML is to teach the algorithm how to learn on its own, store that information, and utilize it to make further inferences. A large enough dataset and the necessary features are important for an ML system to learn.

Deep learning

DL is popularly used in the most predictive data analytics nowadays. The technology's fundamental idea is to develop something resembling the human brain and to train the model to receive, classify, and interpret information in the same way that the brain does, using neural connections, or neural networks.

Federated learning

FL is an arising incipient discipline sometimes known as collaborative learning.[61] It deploys DL or ML algorithms on local datasets present across various decentralized edge devices or servers. The local dataset is not shared as opposed to the centralized ML techniques where all the local datasets are uploaded to a single server. It also differs from traditional decentralized techniques, which assume that the local dataset is uniformly distributed. Figure 17 depicts the fundamental structure of FL.

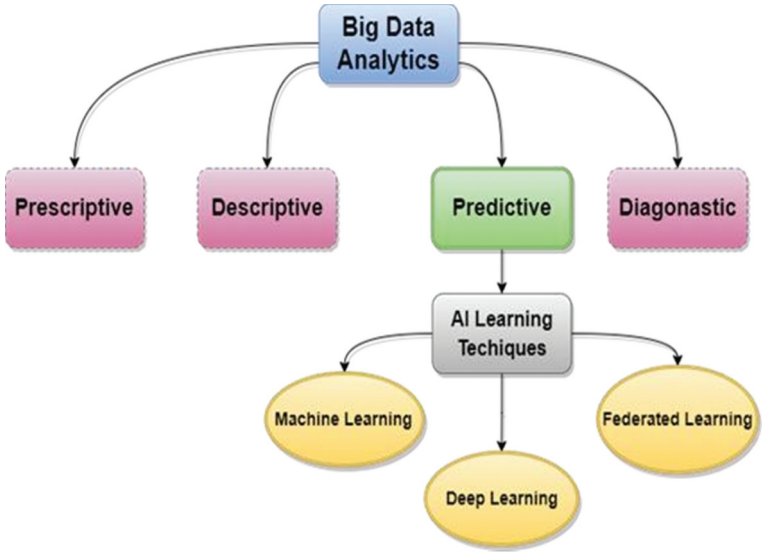

Figure 16. Major learning techniques.

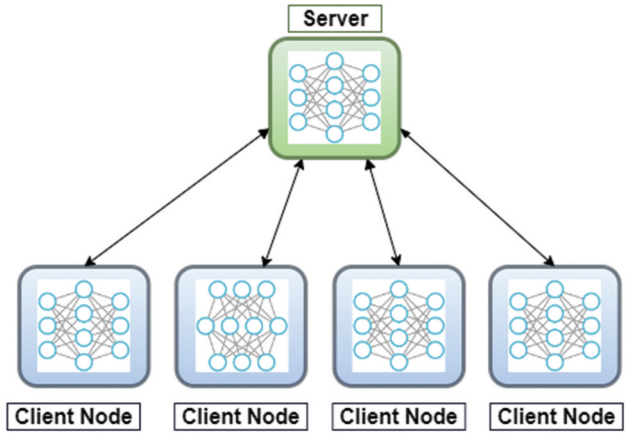

Figure 17. Federated learning.

Upcoming subsections describe how ML, DL, and FL use HE and DP.

Homomorphic encryption

HE is one of the major techniques to provide secure and private learning in the cloud. HE is a kind of encryption that allows third parties to do arithmetic computations directly on ciphertexts without deciphering them. HE is invented by Rivest et al.[65] in 1978. Let us say, m_x and m_y are the plain texts and $E(m_x)$ and $E(m_y)$ are the ciphertexts, then

$$E(m_x) + E(m_y) = E(m_x + m_y) \quad (1)$$

Adding the ciphertexts $E(m_x)$ and $E(m_y)$ is equivalent to the encryption of added plain texts m_x and m_y ($E(m_x+m_y)$) as shown in Equation (1). HE has its own limitations: It works only with integers, ciphertext size increases, while encryption and noise will get increased in ciphertext after each encryption. The initial version of HE was called partial HE that allows only addition or multiplication over ciphertexts. Later, fully homomorphic encryption (FHE) was developed which is capable of supporting more analytics over ciphertexts.[66]

ML algorithms work well with HE in protecting the privacy of training data. The article by Han et al.[67] presented the way to apply logistic regression over homomorphically encrypted data efficiently which preserves the privacy of the learning data. A novel HE framework[68] over non-abelian rings has also been developed and applied over ML training for privacy preservation with better efficiency.

DL also utilizes the HE. DL over homomorphically encrypted data has been experimented with by Hesamifard et al.[69] They encrypted the training data using HE and modified the approximation of activation functions (lowest degree polynomial) of each layer in DL to achieve the privacy of training data. Polynomial approximations are taken from the derivative of each layer's active function methods. The developed model with new approximation methods has been compared with existing learning algorithms over encrypted data. FHE allows more operations to be performed on encrypted data which allows us to apply deep neural network algorithms directly to encrypted data and return encrypted results without compromising security and privacy. Multikey FHE and double encryption with FHE have been introduced with collaborative DL in Li et al.[70] This method provides more privacy of data and the DL model. Privacy-preserving face recognition application has been developed and experimented with the proposed FHE technique.

Not only in DL but in FL also, HE places a major role. In 2020, Madi et al.[62] proposed a model to protect the FL framework with HE and verifiable computing for the Paillier cryptosystem. Subsequently, Zero Knowledge Proof (ZKP) has been added with FL in Guo et al.[71] and Nguyen et al.[72] The best way for a prover to demonstrate to a verifier that a "particular statement is true" while avoiding providing any further information other than the fact that "the statement is true" is through the use of ZKP. Park et al.[63] proposed a PPFL for cloud computing-based service scenarios using individual HE private keys. In FL, instead of training data, only weights of models are encrypted using HE. Weights of models have been aggregated in a centralized server without knowing it which reduces the computational complexity.

Apart from learning techniques, simple data processing algorithms also use HE to provide security. Secure data aggregation technique utilized HE technique with SGX in Silva et al.[73] They have taken smart grid applications and demonstrated secure aggregation with utility preservation and privacy. Based on the analysis of HE, FL makes use of HE's advantages for secure and private learning.

Differential privacy

This section is more focused on DP mechanisms which help to do secure and private learning in the cloud. DP is a privacy-preserving technique introduced by Dwork in 2006.[74] It guarantees that the overall statistics of a dataset remain unchanged despite the change in a single tuple. For example, any algorithm "X" satisfies *ϵ-differential privacy (ϵ-DP)* if it satisfies the following equation 2:

$$Prob[X(D_1) = Ts] \leq exp(\epsilon) Prob[X(D_2) = Ts] \quad (2)$$

where T_s denotes a set of tuples, D_1 and D_2 are any two datasets differing in only a single tuple, and ϵ is the

privacy budget which is an important factor in DP. It lies in the range of 0 (minimum-ϵ) and 1 (maximum-ϵ). The most used standard that can offer a verifiable privacy guarantee is DP.

There has been a lot of research interest in privacy-preserving ML in recent years.[75–77] ML techniques (logistic regression, SVM, etc.) with the DP scheme derive insights from data while protecting individual data privacy. In simple image classification using SVM, the privacy of the training images is protected using ϵ-*DP* with Laplace noise in the article presented by Senekene et al.[78] In clustering, ϵ-*DP* has been utilized by Kaplan et al.[79] and Chaturvedi et al.[80]

A crucial step toward DP is the addition of noise. Abadi et al.[81] added noise in the gradient component of the stochastic gradient descent algorithm to preserve the privacy of training data. Authors trained the model using MNIST with 97% accuracy and CIFAR-10 with 73% accuracy. Subsequently, Nasr et al.[82] presented a simple framework for privacy analysis of DL in 2019 using white-box membership inference attacks then they slightly tuned the performance using denoising in 2020.[83] When it is collaborative learning, privacy becomes important among the participants. This challenge has been resolved by Zhao et al.[84] using ϵ-*DP*. To avoid potential privacy leakage from sharing model parameters, in the training process, the objective function of the neural network was perturbed using a functional mechanism. They proved that the system is robust against unreliable participants.

In FL, Zhao et al.[64] introduced DP while sharing the model parameters with the centralized cloud. Here, the noise is added with model parameters before sending it for aggregation. The same method has been tried to be adapted in cyber-physical systems by Zhang et al.[85] with the added chaotic system. According to the DP study, DL utilizes the advantages of DP in private and secure learning. As previously explained, DP can be used in AC mechanism[24] and secure storage technique[60] also in recent days.

Other techniques

Other than HE and DP, there are other strategies for securing learning processes. This section examines a few of these strategies. Secure multi-party computation (otherwise called secure computation or privacy-preserving computation) is a subdivision of cryptography that seeks to develop methods that allow parties to jointly compute a function over their inputs while keeping those inputs private. Ryffle et al.[86] added MPC with DP in FL to provide a more secure and private FL framework.

BC is also used to preserve the privacy and security of data analytics. In FL, BC has been introduced for secure analytics by Li et al.[87] and Kumar et al.[88] In the recent days, trusted computing has become popular with the greater availability of trusted hardware such as ARM trustzone, Intel SGX, etc. Data analytics in those trusted execution environments are also experimented and verified.[89,90] The main drawback of the trusted hardware is its expensiveness.

There exist different methods to secure the learning techniques such as HE, DP, BC, and Trusted Hardware. Here also, BC and hardware give us secure environment to do analytics whereas HE and DP operate directly on the data before learning them. Hardware solutions are expensive. Data-oriented solutions are fast and cost-effective in the cloud. Table 5 shows an overview of recent secure learning techniques discussed. In general, the performance of the secure and private learning techniques is measured by the model's accuracy after the proposed technique is applied. The secureness of the technique is evaluated by simulating the attacks over the model and by checking its robustness against the attack.

Legend for Table 5 is as follows: **LT**: Learning Technique, **HE**: Homomorphic Encryption, **DP**: Differential Privacy, **Techniques**: Techniques used in secure and private learning technique, **SA**: Whether Security Analysis is done or not? **PA**: Whether Performance Analysis is done or not?

Big data applications - expects secure and private data analytics

In the Big Data era, most cloud applications utilize BDA. **The domains that process or handle user-sensitive data provide more importance to secure and private data analytics in the cloud**. Figure 18 shows the major domains of health care, telecom, finance, and automobile which analyzes more sensitive big data.

A recent article by Ngyan et al.[91] surveyed smart health-care analytics which requires security and emphasizes the importance of private BDA in health care. The finance sector inevitably needs privacy and security when it comes to distributed BDA in the cloud. In the next generation network, IoV plays a major role. PPFL in IoV is captured by Li et al.[92] with practical scenarios. It is clear that whenever user-sensitive data is used for BDA in cloud applications, maintaining the security of the data as well as the privacy of the user is unavoidable.

Table 5. Comparison among the secure and private learning techniques.

LT	Year	Article	HE	DP	Techniques	SA	PA
ML	2019	Han et al.[67]	**Yes**	No	-	Yes	Yes
ML	2020	Li et al.[68]	**Yes**	No	HE over NE rings	Yes	Yes
DL	2017	Hesamifard et al.[69]	**Yes**	No	-	Yes	Yes
DL	2017	Li et al.[70]	**Yes**	No	Multi Key	Yes	Yes
FL	2020	Guo et al.[71]	**Yes**	No	ZKP	Yes	Yes
FL	2021	Madi et al.[62]	**Yes**	No	VC	Yes	Yes
FL	2022	Nguyen et al.[72]	**Yes**	No	ZKP	Yes	Yes
FL	2022	Park et al.[63]	**Yes**	No	Private Keys	Yes	Yes
ML	2018	Kaplan et al.[79]	No	**Yes**	-	Yes	Yes
ML	2019	Senekane et al.[78]	No	**Yes**	-	Yes	Yes
ML	2020	Chaturvedi et al.[80]	No	**Yes**	-	Yes	Yes
DL	2016	Abadi et al.[81]	No	**Yes**	-	Yes	Yes
DL	2019	Nasr et al.[82]	No	**Yes**	-	Yes	Yes
DL	2020	Nasr et al.[83]	No	**Yes**	Denoising	Yes	Yes
DL	2020	Zhao et al.[84]	No	**Yes**	-	Yes	Yes
FL	2021	Zhao et al.[64]	No	**Yes**	-	Yes	Yes
FL	2021	Zhang et al.[85]	No	**Yes**	Chaotic System	Yes	Yes
FL	2018	Ryffel et al.[86]	No	**Yes**	MPC	Yes	Yes
FL	2020	Li et al.[87]	No	No	Blockchain	Yes	Yes
FL	2021	Kumar et al.[88]	No	No	Blockchain	Yes	Yes
FL	2022	Burkhalter et al.[90]	No	No	Intel SGX	Yes	Yes

Figure 18. Major domains.

Open issues and future research directions

After a brief survey on any key research area, it is important to provide future research directions in the research domain. In recent days, BDA has attracted multiple fields where privacy and security are also more concerned. A lot of research work is going on in this big data security and privacy research field. Still, there are opportunities for us to work on this area. In this section, we have discussed some research problems which have to be addressed in this research area and come up with a framework for "secure and private BDA in the cloud" as depicted in Figure 19 for future applications.

Following open research directions could be taken further and develop the research area.

- In a Big Data era, FL is becoming a next-generation learning technique that needs more privacy. In the cloud, Private FL as a service would be a challenging research problem.
- More learning algorithms are running over the cloud where we are relying on mostly third parties. So, Designing trusted learning algorithms should be focused on.
- Hardware security exploitation in BDA in the cloud could be the next challenging research area.
- Developing data-oriented secure and privacy-preserving techniques should be considered.
- Distributed secure storage solutions with dynamic data population.
- Developing auto ML for privacy limitation in sensitive applications.
- In SAC.
  - Combination of techniques can be utilized (e.g.: LDP+BC in RBAC).
  - Quantum-based ideas can be incorporated to improve the security.
  - Single solution for multiple issues such as plain text policy and key abuse can be introduced.
  - Data integrity centric SAC is on demand.
  - Hardware-based SAC could be devised.
- A lot of secure storage mechanisms involve encryption which added an overhead of more execution time and excess storage. So, a memory-efficient and fast secure storage mechanism can be focused on.
- We are moving toward the quantum era. So, developing secure systems based on post-quantum cryptography creates more research problems in that domain.

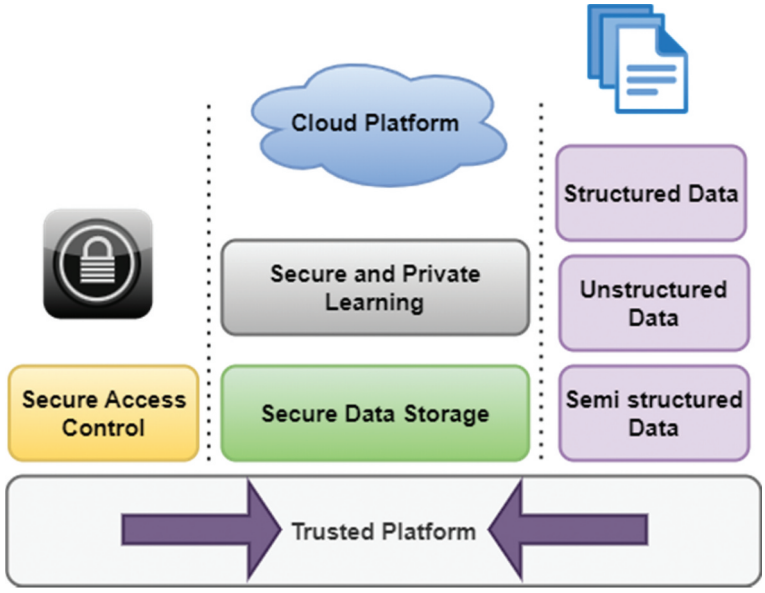

Figure 19. Framework for secure and private BDA.

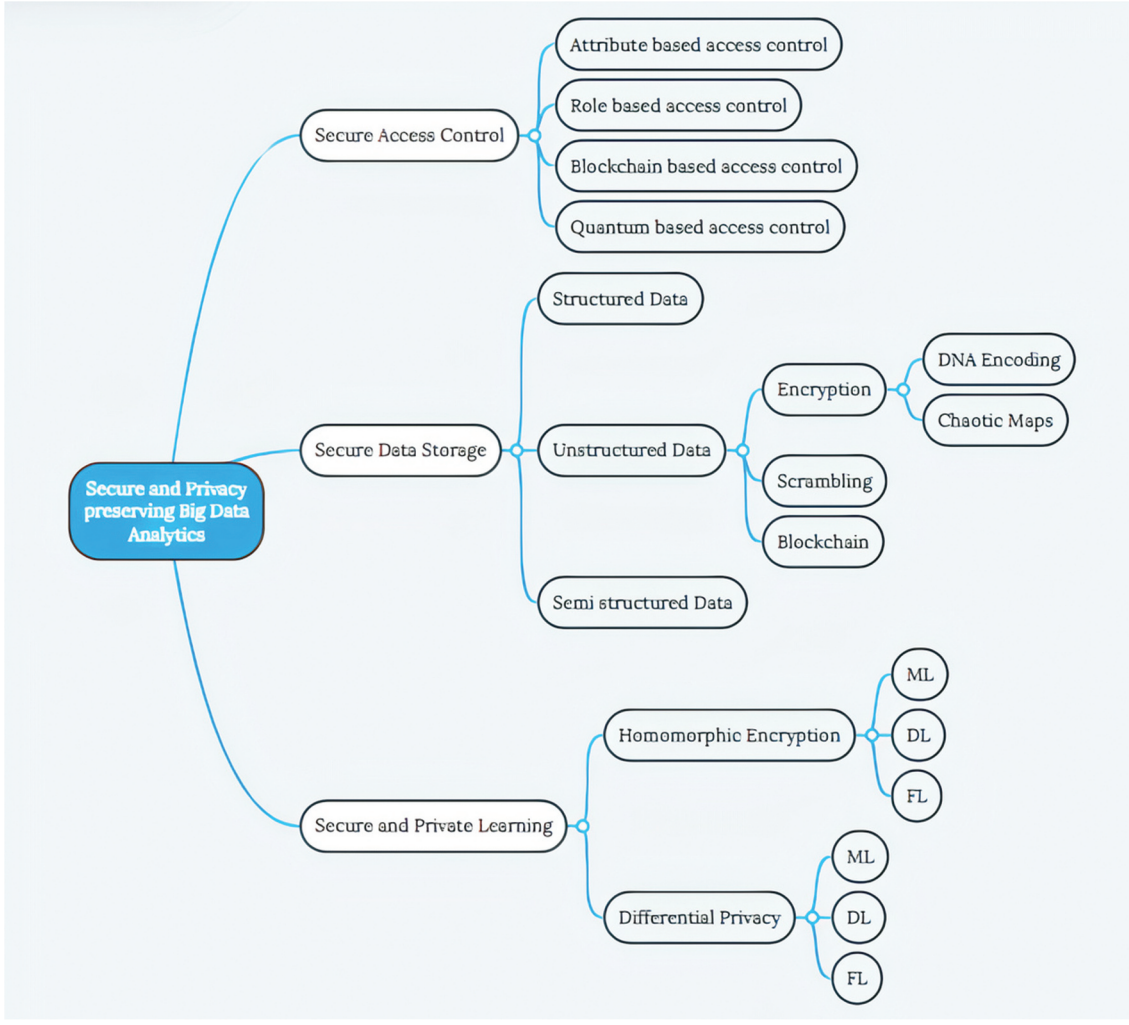

Figure 20. Overall summary of the survey.

Summary and conclusion

In a Big Data era, BDA is in every critical domain which handles sensitive data. It is known that sensitive data should be kept and processed securely as well as privately in the cloud. This article comprehensively surveyed the secure and privacy-preserving BDA in the cloud. It majorly focused on three secure subsystems: SAC, secure data storage, and secure and private BDA. SAC could be provided using role-based, attribute-based, BC-based, and quantum-based mechanisms. Encryption, scrambling, and BC-based approaches are used to secure unstructured data storage. Secure and private learning is achieved through HE and DP. Given the rapid advancement of quantum computing, secure and private BDA may eventually be developed using quantum cryptography techniques in the future. Figure 20 shows the overall summary of the topics discussed in this survey.

Acknowledgment

The authors would like to thankall the reviewers of the article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

No funding was received for this work.

ORCID

Arun Amaithi Rajan http://orcid.org/0000-0002-3019-5879
Vetriselvi V http://orcid.org/0000-0002-3832-6968

References

1. Emilion M. What is big data? [Online]; 2021 [accessed 2023 Jan 28]. https://en.jedha.co/formation-analyse-donnee/big-data.
2. Fang W, Wen XZ, Zheng Y, Zhou M. A survey of big data security and privacy preserving. IETE Tech Rev. 2017;34(5):544–60. doi:10.1080/02564602.2016.1215269.
3. Nti IK, Quarcoo JA, Aning J, Fosu GK. A mini-review of machine learning in big data analytics: applications, challenges, and prospects. Big Data Min Anal. 2022;5(2):81–97. doi:10.26599/BDMA.2021.9020028.
4. Syed D, Zainab A, Ghrayeb A, Refaat SS, Abu-Rub H, Bouhali O. smart grid big data analytics: survey of technologies, techniques, and applications. IEEE Access. 2021;9:59564–85. doi:10.1109/ACCESS.2020.3041178.
5. Pierleoni P, Concetti R, Belli A, Palma L. Amazon, Google and Microsoft Solutions for IoT: architectures and a performance comparison. IEEE Access. 2020;8:5455–70. doi:10.1109/ACCESS.2019.2961511.
6. Sunyaev A. Cloud computing. Internet computing. Vol. 2020. Cham: Springer. doi:10.1007/978-3-030-34957-8_7.
7. Abdulsalam YS, Hedabou M. Security and privacy in cloud computing: technical review. Future Internet. 2022;14(1):11. doi:10.3390/fi14010011.
8. Tran HY, Hu J. Privacy-preserving big data analytics a comprehensive survey. J Pllel Dist Comp. 2019;134:207–18. doi:10.1016/j.jpdc.2019.08.007.
9. El Sibai R, Gemayel N, Bou Abdo J, Demerjian J. A survey on access control mechanisms for cloud computing. Trans Emerg Telecom Tech. 2020;31(2):1–21. doi:10.1002/ett.3720.
10. Pramanik MI, Lau RY, Hossain MS, Rahoman MM, Debnath SK, Rashed MG, Uddin MZ. Privacy preserving big data analytics: a critical analysis of state-of-the-art. Wiley Interdisc Rev: Data Min and Knowl Disc. 2021;11(1). doi:10.1002/widm.1387.
11. Snyder H. Literature review as a research methodology: an overview and guidelines. J Business Research. 2019;104:333–39. doi:10.1016/j.jbusres.2019.07.039.
12. Sun Z, David Strang K, Pambel F. Privacy and security in the big data paradigm. J Comp Infor Syst. 2020;60(2):146–55. doi:10.1080/08874417.2017.1418631.
13. Wu L, Cai HJ, Li H. SGX-UAM: a secure unified access management scheme with one time passwords via intel SGX. IEEE Access. 2021;9:38029–42. doi:10.1109/ACCESS.2021.3063770.
14. Chinnasamy P, Deepalakshmi P, Dutta AK, You J, Joshi GP. Ciphertext-policy attribute- based encryption for cloud storage: toward data privacy and authentication in AI-enabled IoT system. Mathematics. 2022;10(1):68. doi:10.3390/math10010068.
15. Li H, Yang X, Wang H, Wei W, Xue W. A controllable secure blockchain-based electronic healthcare records sharing scheme. J Healthc Eng. 2022;2022:1–11. doi:10.1155/2022/2058497.
16. Akleylek S, Soysaldı M. A new lattice-based authentication scheme for IoT. J Inf Secur Appl. 2022;64(117):103053. doi:https://doi.org/10.1016/j.jisa.2021.103053.
17. Sahi MA, Abbas H, Saleem K, Yang X, Derhab A, Orgun MA, Iqbal W, Rashid I, Yaseen A. Privacy preservation in e-healthcare environments: state of the art and future directions. IEEE Access. 2018;6(4):464–78. doi:10.1109/ACCESS.2017.2767561.
18. Hu VC, Kuhn DR, Ferraiolo DF, Voas J. Attribute-based access control. Computer. 2015;48(2):85–88. doi:10.1109/MC.2015.33.
19. Goyal V, Pandey O, Sahai A, Waters B. Attribute-based encryption for fine-grained access control of encrypted data. 13th Proceedings of the ACM Conference on Computer and Communications Security, 2006, p. 89–98.

20. Bethencourt J, Sahai A, Waters B. Ciphertext-policy attribute-based encryption. IEEE Symposium on Security and Privacy. 2007, p. 321–34.
21. Rasori M, Manna ML, Perazzo P, Dini G. A survey on attribute-based encryption schemes suitable for the internet of things. IEEE Internet Things J. 2022;9 (11):8269–90. doi:10.1109/JIOT.2022.3154039.
22. Hwang YW, Lee IY. A study on CP-ABE-based medical data sharing system with key abuse prevention and verifiable outsourcing in the IoMT environment. Sensors. 2020;20(17):4934. doi:10.3390/s20174934.
23. Wang S, Wang X, Zhang Y. A secure cloud storage framework with access control based on blockchain. IEEE Access. 2019;7:112 713–112 725. doi:10.1109/ACCESS.2019.2929205.
24. Song H, Han X, Lv J, Luo T, Li J. MPLDS: an integration of CP-ABE and local differential privacy for achieving multiple privacy levels data sharing. Peer Peer Netw Appl. 2022;15(1):369–85.
25. Yang Y, Zheng X, Guo W, Liu X, Chang V. Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. Inf Sci (Ny). 2019;479:567–92. doi:10.1016/j.ins.2018.02.005.
26. Chen L, Crampton J. Risk-aware role-based access control. Lecture notes in computer science. Vol. 7170. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-29963-6_11.
27. Alshammari ST, Albeshri A, Alsubhi K. Integrating a high-reliability multicriteria trust evaluation model with task role-based access control for cloud services. Symmetry. 2021;13(3):492. doi:10.3390/sym13030492.
28. Kim J, Park N. Role-based access control video surveillance mechanism modeling in smart contract environment. Trans Emerg Telecommun Technol. 2022;33(4):e4227. doi:10.1002/ett.4227.
29. Nakamoto S, Bitcoin: a peer-to-peer electronic cash system. Cryptography Mailing list, 2009. https://metzdowd.com.
30. Li W, Wu J, Cao J, Chen N, Zhang Q, Buyya R. Blockchain-based trust management in cloud computing systems: a taxonomy, review and future directions. J Cloud Comp. 2021;10(1). doi:10.1186/s13677-021-00247-5.
31. Deep G, Mohana R, Nayyar A, Sanjeevikumar P, Hossain E. Authentication protocol for cloud databases using blockchain mechanism. Sensors (Switzerland). 2019;19(20):10. doi:10.3390/s19204444.
32. Shor PW. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J Comput. 1997;26(5):1484–509. doi:10.1137/S0097539795293172.
33. Fu X, Ding Y, Li H, Ning J, Wu T, Li F. A survey of lattice based expressive attribute based encryption. Comp Sci Rev. 2022;43:100438. doi:10.1016/j.cosrev.2021.100438. [Online]. Available:.
34. Nejatollahi H, Dutt N, Ray S, Regazzoni F, Banerjee I, Cammarota R. Post-quantum lattice-based cryptography implementations: a survey. ACM Comput Surv. 2019;51(6):1–41. doi:https://doi.org/10.1145/3292548.
35. Qiu L, Sun X, Xu J. Categorical quantum cryptography for access control in cloud computing. Soft Comput. 2018;22(19):6363–70. doi:10.1007/s00500-017-2688-2.
36. Huang JJ, Tseng YF, Yang QL, Fan CI. A lattice-based group authentication scheme. Appl Sci (Switzerland). 2018;8(6):1–14. doi:10.3390/app8060987.
37. Reinsel D, Gantz J, Rydning J. Data age 2025: the evolution of data to life-critical, don't focus on big data; focus on the data that's big. White Paper. 2017 Apr. https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf.
38. Cunha M, Mendes R, Vilela JP. A survey of privacy-preserving mechanisms for heterogeneous data types. Comput Sci Rev. 2021;41:100403. [Online]. Available:. doi:10.1016/j.cosrev.2021.100403.
39. Cheng H, Huang Q, Chen F, Wang M, Yan W. Privacy-preserving image watermark embedding method based on edge computing. IEEE Access. 2022;10:18 570–18 582. doi:10.1109/ACCESS.2022.3151115.
40. Moosavi SR, Izadifar A. End-to-end security scheme for e-health systems using DNA-based ecc. Silicon Valley Cybersecurity Conference, 2022. p. 77–89.
41. Li S, Chen X, Wang Z, Qian Z, Zhang X. Data hiding in iris image for privacy protection. IETE Tech Rev. 2018;35(sup1):34–41. doi:10.1080/02564602.2018.1520153.
42. Carvalho T, Moniz N, Faria P, Antunes L. Survey on privacy-preserving techniques for data publishing. 2022;1(1) http://arxiv.org/abs/2201.08120.
43. Kaur M, Kumar V. A comprehensive review on image encryption techniques. Arch Comput Methods Eng. 2020;27(1):15–43. doi:10.1007/s11831-018-9298-8.
44. Zia U, McCartney M, Scotney B, Martinez J, AbuTair M, Memon J, Sajjad A. Survey on image encryption techniques using chaotic maps in spatial, transform and spatiotemporal domains. Int J Inf Secur. 2022;21(4):917–35. doi:10.1007/s10207-022-00588-5.
45. Abduljabbar ZA, Abduljaleel IQ, Ma J, Sibahee MAA, Nyangaresi VO, Honi DG, Abdulsada AI, Jiao X. Provably secure and fast color image encryption algorithm based on s-boxes and hyperchaotic map. IEEE Access. 2022;10:26 257–26 270. doi:10.1109/ACCESS.2022.3151174.
46. Pan H, Lei Y, Jian C. Research on digital image encryption algorithm based on double logistic chaotic map. EURASIP J Image and Video Process. 2018;2018(1). doi:10.1186/s13640-018-0386-3.
47. Wan Y, Gu S, Du B. A new image encryption algorithm based on composite chaos and hyperchaos combined with DNA coding. Entropy. 2020;22(2):171. doi:10.3390/e22020171.
48. Huang H, Cheng D. 3-image bit-level encryption algorithm based on 3D nonequilateral Arnold transformation and hyperchaotic system. Secur Commun Netw. 2020;2020:1–13. doi:10.1155/2020/8841302.
49. Ahmad I, Shin S. Encryption-then-compression system for cloud-based medical image services. 2022

International Conference on Information Networking (ICOIN), 2022, p. 30–33.
50. Namasudra S. A secure cryptosystem using DNA cryptography and DNA steganography for the cloud-based IoT infrastructure. Comput Electr Eng. 2022;104:12. doi:10.1016/j.compeleceng.2022.108426.
51. Bao W, Zhu C. A secure and robust image encryption algorithm based on compressive sensing and DNA coding. Multimed Tools Appl. 2022;81(11):15977–96. doi:10.1007/s11042-022-12623-7.
52. Choi J, Yu NY. Secure image encryption based on compressed sensing and scrambling for internet-of-multimedia things. IEEE Access. 2022;10:10 706– 10 718. doi:10.1109/ACCESS.2022.3145005.
53. Arora A, Sharma RK. Cryptanalysis and enhancement of image encryption scheme based on word-oriented feedback shift register. Multimed Tools Appl. 2022;81(12):16 679–16 705. doi:10.1007/s11042-022-11973-6.
54. Sukumar A, Subramaniyaswamy V, Ravi L, Vijayakumar V, Indragandhi V. Robust image steganography approach based on RIWT-Laplacian pyramid and histogram shifting using deep learning. Multimed Syst. 2021;27(4):651–66. doi:10.1007/s00530-020-00665-6.
55. Xia Z, Wang L, Tang J, Xiong NN, Weng J. A privacy-preserving image retrieval scheme using secure local binary pattern in cloud computing. IEEE Trans Netw Sci Eng. 2021;8(1):318–30. doi:10.1109/TNSE.2020.3038218.
56. Ali A, Khan A, Ahmed M, Jeon G. Bcals: blockchain-based secure log management system for cloud computing. Transactions on Emerging Telecommunications Technologies, 2021.
57. Vetriselvi V, Pragatheeswaran S, Thirunavukkarasu V, Arun AR. Preventing forgeries by securing healthcare data using blockchain technology. In: Information and communication technology for sustainable development. Springer Singapore; 2020. pp. 151–59.
58. Alqaralleh BA, Vaiyapuri T, Parvathy VS, Gupta D, Khanna A, Shankar K. Blockchain- assisted secure image transmission and diagnosis model on internet of medical things environment. Personal and Ubiquitous Computing, 2021.
59. Alhazmi HE, Eassa FE, Sandokji SM. Towards big data security framework by leveraging fragmentation and blockchain technology. IEEE Access. 2022;10:10 768–10 782. doi:10.1109/ACCESS.2022.3144632.
60. Liu B, Ding M, Xue H, Zhu T, Ye D, Song L, Zhou W. DP-image: differential privacy for image data in feature space. 2021. http://arxiv.org/abs/2103.07073
61. Yin X, Zhu Y, Hu J. A comprehensive survey of privacy-preserving federated learning: a taxonomy, Review, and Future Directions. ACM Comput Surv. 2021;54(6):1–36. doi:10.1145/3460427.
62. Madi A, Stan O, Mayoue A, Grivet-S´ebert A, Gouy-Pailler C, Sirdey R. A secure federated learning framework using homomorphic encryption and verifiable computing. In: 2021 reconciling data analytics, automation, privacy, and security: a big data challenge (RDAAPS). 2021. pp. 1–8.
63. Park J, Lim H. Privacy-preserving federated learning using homomorphic encryption. Appl Sci (Switzerland). 2022;12(2):734. doi:10.3390/app12020734.
64. Zhao B, Fan K, Yang K, Wang Z, Li H, Yang Y. Anonymous and privacy-preserving federated learning with industrial big data. IEEE Trans Industr Inform. 2021;17(9):6314–23. doi:10.1109/TII.2021.3052183.
65. Rivest RL, Adleman L, Dertouzos ML. On data banks and privacy homomorphisms. Foundations of secure computation. Academia Press; 1978. pp. 169–79.
66. Munjal K, Bhatia RA. Systematic review of homomorphic encryption and its contributions in healthcare industry. Complex Intell Syst. 2022. doi:10.1007/s40747-022-00756-z.
67. Han K, Hong S, Cheon JH, Park D. Logistic regression on homomorphic encrypted data at scale. 33rd AAAI Conference on Artificial Intelligence, 2019, p. 9466–71.
68. Li J, Kuang X, Lin S, Ma X, Tang Y. Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. Inf Sci (Ny). 2020;526:166–79. doi:10.1016/j.ins.2020.03.041.
69. Hesamifard E, Takabi H, Ghasemi M. CryptoDL: deep neural networks over encrypted data. pp. 1–21, 2017. http://arxiv.org/abs/1711.05189
70. Li P, Li J, Huang Z, Li T, Gao CZ, Yiu SM, Chen K. Multi-key privacy- preserving deep learning in cloud computing. Future Gener Comput Syst. 2017;74:76–85. doi:10.1016/j.future.2017.02.006.
71. Guo J, Liu Z, Lam K-Y, Zhao J, Chen Y, Xing C. Secure weighted aggregation for federated learning, p. 1–18, 2020. http://arxiv.org/abs/2010.08730.
72. Nguyen T, Thai M. Preserving privacy and security in federated learning, 2022. https://arxiv.org/abs/2202.03402.
73. Silva LV, Barbosa P, Marinho R, Brito A. Security and privacy aware data aggregation on cloud computing. J Internet Ser Appl. 2018;9(1). doi:10.1186/s13174-018-0078-3.
74. Dwork C. Differential privacy. In: Automata, languages and programming. Berlin Heidelberg: Springer; 2006. pp. 1–12.
75. Al-Rubaie M, Chang JM. Privacy-preserving machine learning: threats and solutions. IEEE Secur Priv. 2019;17(2):49–58. doi:10.1109/MSEC.2018.2888775.
76. Li X, He J, Vijayakumar P, Zhang X, Chang V. A verifiable privacy-preserving machine learning prediction scheme for edge-enhanced hcpss. IEEE Trans Industr Inform. 2022;18(8):5494–503. doi:10.1109/TII.2021.3110808.
77. Panzade P, Takabi D. Towards faster functional encryption for privacy-preserving machine learning. 2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 2021, p. 21–30.
78. Senekane M. Differentially private image classification using support vector machine and differential privacy. Mach Learn Knowl Extr. 2019;1(1):483–91. doi:10.3390/make1010029.
79. Kaplan H, Stemmer U. Differentially private k-means with constant multiplicative error. Adv Neural Inf Process Syst. 2018;5431–41.

80. Chaturvedi A, Nguyen H, Xu E. Differentially private k-means clustering via exponential mechanism and max cover, 2020. http://arxiv.org/abs/2009.01220
81. Abadi M, McMahan HB, Chu A, Mironov I, Zhang L, Goodfellow I, Talwar K. Deep learning with differential privacy. Proceedings of the ACM Conference on Computer and Communications Security, 2016, p. 308–18.
82. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning. 2019 IEEE Symposium on Security and Privacy, 2019, p. 739–53.
83. Nasr M, Shokri R, Houmansadr A. Improving deep learning with differential privacy using gradient encoding and denoising, p. 1–15, 2020. [Online]: http://arxiv.org/abs/2007.11524
84. Zhao L, Wang Q, Zou Q, Zhang Y, Chen Y. Privacy-preserving collaborative deep learning with unreliable participants. IEEE Trans Inf Forensics Secur. 2020;15:1486–500. doi:10.1109/TIFS.2019.2939713.
85. Zhang Z, Zhang L, Li Q, Wang K, He N, Gao T. Privacy-enhanced momentum federated learning via differential privacy and chaotic system in industrial Cyber–Physical systems. ISA Trans. 2021;128:17–31. doi:10.1016/j.isatra.2021.09.007.
86. Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D, Passerat-Palmbach J. A generic framework for privacy preserving deep learning, pp. 1–5, 2018. [Online]. Available: http://arxiv.org/abs/1811.04017
87. Li Z, Liu J, Hao J, Wang H, Xian M. CrowdSFL: a secure crowd computing framework based on blockchain and federated learning. Electronics (Switzerland). 2020;9(5):773. doi:10.3390/electronics9050773.
88. Kumar R, Khan AA, Zakria JK, Golilarz NA, Zhang Y, Ting S, Zheng W, Wang C. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. IEEE Sens J. 2021;21(14). 16 301–16 314. doi:10.1109/JSEN.2021.3076767.
89. Chandra S, Karande V, Lin Z, Khan L, Kantarcioglu M, Thuraisingham B. Securing data analytics on SGX with randomization, lecture notes in computer science. Vol. 10492, LNCS; 2017. pp. 352–69.
90. Burkhalter L, Lycklama H, Viand A, Kuchler N, Hithnawi A. RoFL: attestable robustness for secure federated learning. 2021. [Online]. http://arxiv.org/abs/2107.03311.
91. Nguyen DC, Pham Q-V, Pathirana PN, Ding M, Seneviratne A, Lin Z, Dobre O, Hwang W-J. Federated learning for smart healthcare: a survey. ACM Comput Surv. 2022;55(3):1–37. doi:10.1145/3501296.
92. Li Y, Li H, Xu G, Xiang T, Lu R. Practical privacy-preserving federated learning in vehicular fog computing. IEEE Trans Veh Technol. 2022;9545:1–14.