Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 4586-4616
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.472

Security Issues And Challenges Of Big Data Over Clouds: A Survey

Lalit Mohan Gupta Department of Computer Science Engineering, Aligarh College of Technology and Management, Aligarh
Email: lalitmguptaamu@gmail.com

Abdus Samad Department of Computer Engg., Zakir Husain College of Engg.& Tech. Aligarh Muslim University,Aligarh E-mail:abdussamadamu@gmail.com

Hitendra Garg Department of Computer Engineering and Applications, GLA University, Mathura, Uttar Pradesh, India E-mail: hitendra.garg@gmail.com

Abstract: The inherent dynamic features of data offer huge concerns and hurdles for researchers when it comes to storing, processing, and interpreting big data. Due to its massive, varied, and complicated data sets, it necessitates a lot of storage and processing power. Distributed systems are commonly used to meet these criteria. Big data has been implemented in the cloud environment for the goal of cost-cutting and easy management. Since cloud computing encourages virtualization, protecting one's data becomes a more difficult task. Several privacy-preserving strategies have been created at various stages of the big data life cycle to acquire additional security and privacy over the data. However, existing privacy-preserving technologies are insufficient to provide complete data protection. As a result, the goal of this article is to provide a general review of the advantages and pitfalls of several state-of-the-art privacy protection strategies for big data in the cloud. This work, in particular, demonstrates a thorough examination of existing methodologies' effectiveness and usefulness in the current situation. Various possible remedies are also proposed based on the research. The work also tackles future big data research aspects relating to privacy preservation in cloud computing contexts.

Keywords: Attribute based encryption, Identity based encryption, Fully homomorphic encryption, Security, Privacy

1. Introduction

Big data has gained the most popularity among researchers, academicians, corporations, and scientists as the demand for the internet continues to grow around the world. The reason for its increasing prominence in enterprises and corporations is that they have amassed a massive amount of data that did not exist just a few years ago. Big data is generated in large quantities from a variety of sources that it becomes very difficult to manage such massive amounts of data. Data analysis and processing are beyond the capabilities of traditional computing systems. As a result, we link them to different tools and procedures. It should be mentioned that 90% of the data has

been generated in the last few years. An standard relational database system is unsuitable for storing this information. As a result, we classify them as distinct types of data structures and databases. One approach for dealing with complex and massive data sets is to move them to the cloud. Data deployment on the cloud provides the most flexibility to the data owner. In the global internet market, more flexibility means more obstacles. In the age of big data analytics, good data management has become a necessary part for managing vast amounts of data. The widespread usage of cloud computing has resulted in a massive increase in data volumes and also diverse formats of data. To deal with such a wide range of data formats, efficient management necessitated addressing numerous concerns such as data size, variety, security, and so on.

2. Big Data Overview

In terms of vast size, complexity, and management, the development of massive amounts of data format from various data sources and collected into the storage device becomes a big difficulty for researchers. Big data has offered a variety of frameworks to address these issues. Until today, there has been no clear definition of big data. Different researchers have described it in their own distinct ways. Big data is defined as "any collection of massive and complicated datasets that cannot be analyzed using typical computing techniques," according to Wikipedia. IDC provides a common definition of big data: "Big data technologies" refers to a new generation of technologies and architectures aimed at extracting value from massive amounts of data from a range of sources at a low cost, by allowing the collection, analysis, and visualization of vast amounts of data [1]. Big data is no longer merely a collection of data; it has evolved into a full-fledged subject for academics and researchers, requiring the development and study of a variety of tools, approaches, and frameworks. Every day, new technologies, devices, and communications are developed. As a result, the amount of huge data produced by humanity is rapidly increasing. With the use of recent technological developments such as social media, the Internet of Things (IoT), sensor networks, healthcare applications, multi-media, cell phones, military websites, and many other businesses, a massive amount of data is being generated that is rapidly increasing around the world. Figure 1 depicts some of the primary fields that fall under the umbrella of big data.

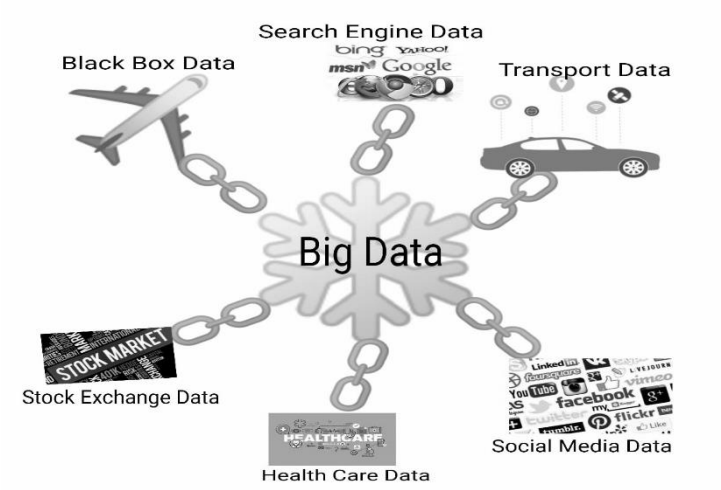

Fig. 1.Major sources of big data

• **Black Box Data:** Data stored in the black box of aeroplanes, helicopters, and jets: The black box of aeroplanes, helicopters, and jets is used to record microphone voices, performance information on aeroplanes, and so on.

• **Data from social media sites:** Various social media sites save information about individuals all over the world.

• **Data from the stock exchange:** It contains information on stock purchases and sales, among other things.

• **Transportation data:** Transportation data contains information about a vehicle's model, capacity, distance, and many other characteristics.

• **Data from search engines:** Different search engines retrieve information from various databases.

• **Healthcare data:** Big data is data that comes from a variety of sources in diverse formats such as text, audio, image, video, and so on, in vast quantities at a high rate. The created data could be in a variety of formats, such as structured, semi-structured, or unstructured, depending on the data type. Because the rate at which data is generated is rapidly rising each second, storing and managing data using traditional methods is becoming increasingly difficult [2]. Using the increasing variety of data, handling huge and complicated data sets with traditional methods becomes increasingly difficult. When it comes to maintaining data warehouses and processing data, big data presents a number of issues. As a result, an efficient system must be developed that can collect vast amounts of complicated data and analyse it effectively. Big data has broadened the field of scientific study by changing traditional commercial models into scientific values, and it has the potential to strengthen the economy [3]. Businesses and organizations can also employ data analysis to extract useful information from large amounts of data. Data analysts can use data analysis tools to improve the capacity of internal decision making, make inferences, and create new opportunities. The properties of big data are defined by the three V's: volume,

velocity, and variety. Deep research reveals that the three V's definition alone is insufficient to describe large data. In addition, the usefulness and validity of big data are included to provide a more comprehensive explanation of big data. Figure 2 depicts the five primary features of big data: volume, velocity, diversity, value, and validity.

Fig.2 Illustration of5V's of Big Data

Volume: Volume: This phrase refers to all forms of data that have been acquired in large quantities by various resources and that are continuing to grow in size. The key benefit is that it allows you to collect a large amount of data in order to uncover useful hidden records and patterns through data analysis. Data size is rapidly growing as the use of social networking, e-commerce, and e-governance sites grows at a rapid pace. Laurila et al. [4] devised a mechanism for collecting unique longitudinal data from digitalizing smartphone devices and handing it over to a study group. Every second, Google receives 43000 queries, Twitter receives 7000 previous tweets, and 24 lacs of emails are exchanged. YouTube has 80K video views and generates 21 TB of internet traffic. By the year 2020, one-third of the data will be stored in the cloud or to be transferred through the cloud.

Velocity: The term velocity refers to the rate at which new data is generated from various sources and processed. The fundamental characteristic of velocity is how quickly new data must be processed. In another sense, velocity can be thought of as the rate at which data is transferred from one entity to another. Because of the absorption of supplementary data collections, the summary of previously preserved data or legacy collections, and streamed data that arrives from numerous sources, the contents and data size are constantly changing [5].

Variety: One of the most important characteristics of big data is its diversity, the data can be form of text, audio, log files, images, or videos. Variety is a term used to describe the property of data. Alternatively, the phrase variety refers to the various formats of data collected through the health-care system, digitalize sensors and smart phones, or social networking sites. A relational database, movies, photos, texts, audios, and data logs are examples of several data types that can be classified as structured or unstructured. Through mobile devices and sensors, Internet users

generate massive amounts of structured data [6] such as relational databases, online games, text messages, blogs, and social media content generate many sorts of unstructured data. A relational database is a type of organised data that stores information in tabular form, i.e. rows and columns. The majority of data generated by mobile applications is unstructured.

Value: If someone is thrilled to extract some valuable information from an existing database, value may play a vital function in the field of Big data. To put it another way, it is the process of extracting vast amounts of hidden meaningful values from a big number of datasets of various types [7].

Validity: The term validity refers to the accuracy and correctness of data for the intended application. When data is accurate and reliable, it aids business analysts in making the best judgments possible.

Big data may be effectively utilised and used to forge new paths in a variety of human undertakings. It aids humans in gaining a greater understanding of the world. The growing amount of data, on the other hand, is posing a threat to user privacy. Various e-commerce sites, including as Flipkart, Amazon, Paytm, and Google, can, for example, analyse our purchase preferences and browsing habits based on our previous search history. Similarly, social networking services (such as Twitter, Instagram, and Facebook) store all of their members' sensitive information. Youtube, one of the most prominent video-sharing websites, can predict our video preferences and search habits.

3. Privacy Issues in Big data

For researchers, the security and privacy of big data has become a serious concern. Because of advances in technology in the areas of analytical tools and knowledge extraction applications, examining data and extracting usable information or patterns from enormous data sets has become more easier. These technologies may reveal personal information about the user that the user does not wish to be made public. As big data is increasingly being used to acquire, retain, and reuse personal information in order to gain commercial advantage, it may pose a threat to one's privacy and security. The following conditions [8] may constitute a breach of user confidentiality:

- By combining one's data with existing storage databases, additional information about users can be inferred. This information may be private and personal to those who do not wish for it to be made public.
- If confidential data are obtained and kept in an unsecured place, processing these data from an unsecured location may result in data leaking.

Data production, data gathering, and data processing are the three primary stages of big data development. Researchers have created different privacy strategies based on various aspects of big data development in recent years to address privacy for big data. To keep data private, falsifying data techniques and access limitations are utilised during the data production process. In access restriction techniques, data

owners specify the data access limit so that only selected users have access to individuals' private data, but in falsifying techniques, original data is modified before transmission to a third party who may or may not be trustworthy. Encryption is the most common method for achieving privacy in big data throughout the data storage phase. Encryption is the process of converting a plaintext communication into a secret code using a specific method known as ciphertext. The adversary should not be able to decipher this ciphertext readily. The more difficult it is to decipher the meaning of ciphertexts, the more secure the data is. To achieve perfect data security, the adversary's understanding of ciphertext should be low. The researchers have introduced various well-known encryption-based strategies to address data security. Identity-based Encryption (IBE), Attribute-based Encryption (ABE), and Storage Path Encryption are the three kinds in general. These encryption algorithms are excellent for the security of cloud-based outsourced data. The sensitiveness of the information, or how much information is sensitive, determines cloud security. One simple way to keep data safe in the cloud is to send sensitive data to a private cloud and non-sensitive data to a public cloud. However, because the number of users (authorised or unauthorised) varies based on the services, this technique fails due to limitations in data accessibility among data users. As a result, sensitive data requires robust privacy-preserving measures to ensure that the privacy and security of such sensitive data is maintained at a high degree. There are several data privacy-preserving processes and knowledge extraction methods for extracting and learning important information from stored data that are widely used throughout data processing. Some anonymization techniques, such as generalisation and suppression, can be used to protect data privacy. The primary goal of data processing is to assure the data's effectiveness while maintaining privacy. Knowledge extraction techniques like clustering, classification, and association rule mining, on the other hand, are used to extract useful records and patterns from huge and complex datasets. Clustering and classification algorithms can be used to partition input data into distinct groups, while association rule mining can be used to uncover valuable associations and fashions in the input data. Despite the fact that many scholars have published multiple research articles on big data, only a few of them write survey/review papers [3],[9]. These publications do a decent job of describing the fundamental concepts of confidentiality, security, and privacy preservation in large data, but they fall short of covering all facets of the subject. Although there are various important approaches to this topic, just a few are discussed in the research paper. Researchers in [3],[9], for example, did not succeed in presenting complete discussions of many elements and methods to big data privacy for cloud computing. There are no prospective issues in this area as well.

Infrastructure of Big Data

Big data is a collection of different data dimensions that enter at a high rate from various data sources. To properly manage a variety of data, an effective and powerful framework must be designed to process a huge amount of data received quickly from many sources. There are several stages to the Big Data life cycle. As demonstrated in Fig. 3, this involves data creation, data storage, and data processing. Traditionally, all generated data from diverse data sources is stored and processed in a single storage

system. The amount, pace, diversity, value, and veracity of big data may all be examined. However, current technology employs the concept of data dispersion to improve scalability, accessibility, and computing time. A single application may use several data storage to fetch the relevant information in data dissemination. Because data is now stored and processed in a distributed method. Big data storage and processing are investigated using the most popular cloud computing technologies, such as Hadoop MapReduce.

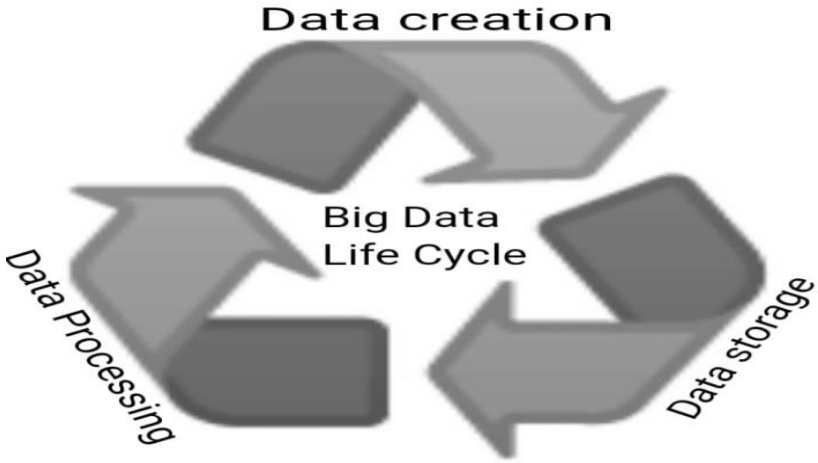

Fig.3 Big Data Life Cycle

This section describes the life cycle of big data. Moreover, a detailed discussion on the popularity of the deployment of big data on cloud, cloud computing technology and its challenges is carried out.

4.1 Life Cycle Of Big Data

- **Data creation:** During the data creation phase, a large amount of data is collected from many distributed sources such as sensors, telephones, transaction logs, social networking sites, and so on. Since the previous few years, the amount of data produced by humans and robots has increased tremendously. Instagram, for example, employs Amazon Web Services (AWS) to store and manage a vast volume of data created by Instagram users, such as over 35 snaps that receive close to 150 likes in a second. Every day, 6000 billion business transactions are scrutinised in the hopes of uncovering fraud. For example, social networking sites like Facebook contribute significantly to the daily production of 35TB of new data. According to Forbes (http://www.forbes.com/), public cloud services have become one of the most popular among cloud customers around the world. The data is usually gathered from a variety of sources and stored in a centralised storage system. This storage system contains a large number of different data formats, such as structured, semi-structured, and unstructured data. As a result, traditional systems find it challenging to handle them effectively.
- **Data storage:** The word "data storage" refers to the collection and management of massive volumes of data created by a variety of sources. Hardware

infrastructure and data management are the two fundamental components of a data storage system. The hardware infrastructure is in charge of putting all of the collected information and communications technology resources to use in order to complete various tasks. Data management [10] is a tool that consists of a collection of software that runs on top of hardware. It provides a variety of interfaces for interacting with and analysing stored data. These technologies are used to handle and query massive amounts of data.

- **Data processing:** After storing a large amount of data in a storage system. The term "data process" refers to the transformation of data, as well as the collection, pre-processing, and extraction of useful information. Data storage is essential since data can come from a variety of sources, such as websites with text, images, and videos.

4.2 Big data and Cloud Computing

Big data requires a lot of processing and storage power, which can be met by using cloud computing technology. Cloud computing provides firms, organisations, and businesses with a variety of benefits, including quick processing power, on-demand storage capacity, scalability, and easy data access. They frequently use the cloud to take advantage of its various services. Virtualization, distributed storage, and processing are the most prominent cloud computing aspects, making it a more powerful technology in today's world. Cloud computing is more efficient and quick in performing computation operations on huge and complex data in a way that was previously difficult. Despite this, most businesses are hesitant to transfer sensitive or secret data to the cloud due to concerns about data privacy and security. They refused to outsource their data unless they were certain that it would be entirely secure in the cloud. Figure 4 depicts the cloud's basic design. Data owner, data storage server, and crypto-system are the three basic components of a cloud environment. It is the responsibility of the data owner to outsource his data to the cloud. The data storage server stores all data that the data owner outsources. To be considered a non-trustworthy entity, crypto-systems conduct compute activities. After outsourcing data to the cloud, the data owner loses physical control of the data. A crypto-system is in charge of doing all cloud computations, and an untrusted entity may leak secret information. As a result, cloud computing raises major privacy concerns.

- **Outsourcing**: Data owners use the cloud to store their information. This way, data owners not only save storage costs and computational costs, but also increase the level of accessible to authorised users. However, outsourcing data to the cloud by the data owner results in a loss of physical control over his data, which is one of the leading causes of cloud insecurity: data on the cloud can be leaked or damaged by someone for the purpose of gaining a competitive advantage or engaging in malicious activity. To address these concerns, a secure computing method and data storage system must be chosen. The verification integrity plays a critical role in ensuring data privacy for cloud users on outsourced data, ensuring that his data is completely stored in the cloud. Furthermore, an effective system must be developed to allow data owners to check data integrity and confidentiality.

- **Multi-tenancy**: Virtualization is another term for multi-tenancy. Virtualization allows numerous users to share the same cloud platform. With the help of a resource allocation policy, data from various cloud users is stored on identical physical data storage. In this circumstance, a cloud provider considers that all cloud users are completely trustworthy and honest, yet another cloud user who is already a customer of the same cloud provider may pose a threat. Another risk is that a user who is not a current subscriber of the cloud provider could compromise the system's privacy by stealing someone's data. As a result, such activities raise the risk of data and computation leakage for researchers. As a result, developing a process that maintains data privacy and security is critical.
- **Massive computation:** Traditional procedures are inadequate for dealing with large amounts of data storage and complex data computation while maintaining individual privacy. As a result, a system that can efficiently manage data storage and calculation tasks is required.

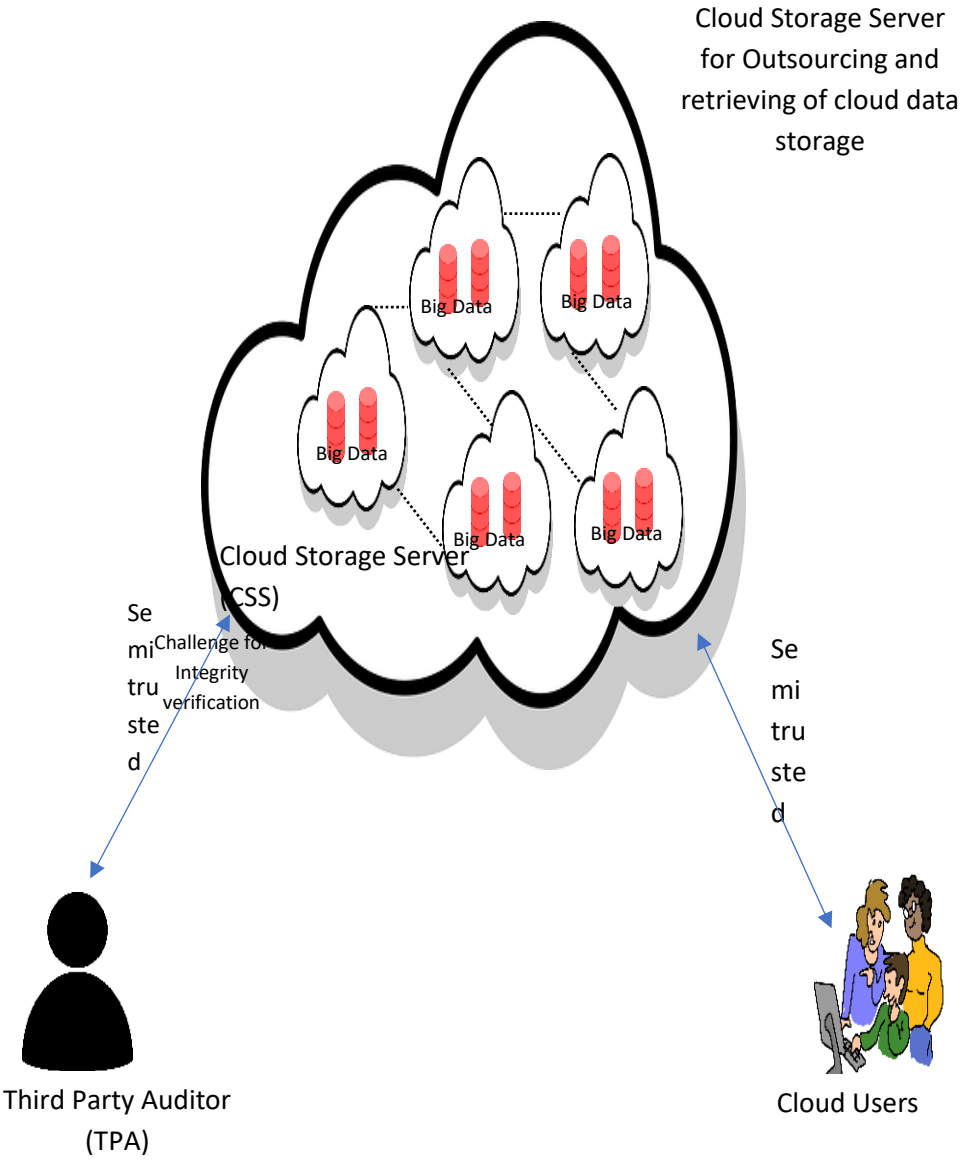

Fig.1 Architecture of Cloud Computing

4. Various Privacy / Confidential Issues in Data Generation Phase

Active and passive data are two types of data generated from diverse sources. Active data is that which the data owner desires to make available to a third party, whereas passive data is that which is gathered by the third party through the data owner's online behaviour, such as browsing. The data owner may be completely ignorant that their information is being collected by a third party. As a result, it is vital for data owners to maintain data privacy during browsing activities and to safeguard their data from being stored by a third party. A data owner's need to keep all of his sensitive

and personal information hidden as much as possible. Restricting access policies or faking data techniques can be used to mitigate the impact of attacks on individual privacy during data generation [9].

5.1. **Access Restriction**

One of the most popular features of cloud computing is data restriction, which allows the data owner to manage the accessibility of certain data to specific users after outsourcing his data to the cloud. This feature includes the ability for the data owner to distribute only those elements of the data that he desires. If he believes that specific material may expose his sensitive data, he simply refuses to access it. To maintain data privacy in a cloud setting, the data owner must adopt an effective fine-grained access control system and ensure that data has not been taken secretly by any permitted or unauthorised party. If the data owner wishes to transmit the data passively to a third party, he must use encryption tools [9], script blockers, and anti-tracking extensions tools to ensure that the data remains private and that access to the confidential data is limited. There are also some other tools, such as anti-software and anti-malware software, that can be used to maintain data confidentiality on his laptop or computer.

5.2. **Fabricating Data**

Because access restriction does not offer a complete solution for data privacy, data owners may fail to keep their data confidential while using access restriction policies. Distort the data using certain software tools before the data is received by the third party in specified scenarios. It becomes extremely difficult for the opponent to recover the true information if the data has been tampered with. As a result, the data privacy is preserved by the distortion techniques. The data owners utilised the following tools to fake the data in order to hide an individual's online identity [9].

- **Socketpuppet** is a software security tool used by data owners to hide online characteristics of individuals through a variety of methods. Individual online activities are concealed by assuming a false identity and posing as someone else, so that information belonging to one person appears to belong to someone else. In this method, data owners are able to conceal individual identities from data collectors who are unable to obtain sufficient information about a single person. As a result, people are unable to identify the user's genuine identity, and personal information cannot be easily divulged.
- **Maskme** is another security tool that is designed to conceal an individual's genuine online identity. This technique is especially useful when the data owner needs to reveal debit or credit card information when making an online purchase. The user can generate the assumed name of his private data, such as debit/credit card numbers and email addresses, using this programme. The data owner can use these assumed names whenever he wants.

5.3. **Privacy In Data Storage Phase**

With recent advancements in data storage technologies, such as the rise of cloud computing, storing large volumes of data sets is no longer a problem, however data security is a key concern for academics and researchers. If numerous users have access to a single data storage unit, the security of the data may be jeopardised, and the disclosure of specific sensitive information can be extremely damaging. As a result, we needed to create a system that protects data privacy from the attacker. In the traditional model, data centres are responsible for performing complicated computations, data interchange, and data retrieval. To solve long processing times, distributed systems are introduced, in which a single application may demand data sets from multiple data centres, posing a data security concern. Application-level encryption [12], media-level security, database-level security, and file-level data security are the four levels of current data security technologies. The most challenging study topic has been safeguarding data security and privacy [13], [14], [15] for present storage architectures [16], however these solutions fail to support the big data analytics platform. The storage infrastructure should be scalable and dynamically configurable to support a variety of applications. One of the most important mechanisms for concentrating the requirements of the future cloud computing paradigm is storage virtualization [17]. In storage virtualization, several network storage devices are connected in a way that seems to be a single storage device.

5. Several Approches to Privacy Preservation

6.1 Storage Cloud

When a large amount of data is generated by various resources, it is necessary to store it on cloud servers abroad. During the data storage phase, the cloud server is regarded to be an untrustworthy entity that must manage all data activities. Because cloud servers store all of the data, they may be compelled to expose all of it to competitors or marketing teams in exchange for a monetary reward. As a result, before storing data in the cloud, the data owner must implement a specialised security method to ensure data protection. Data security is primarily three-dimensional. The first is confidentiality, in which data is kept private and only the authorised user is aware of it. The second is integrity, which ensures that only authorised users have the capacity to modify data, and the third is availability [11], which ensures that data is always available to authorised users. Because these terms are directly related to data privacy, any violation of data confidentiality or integrity by anybody will have a direct influence on the privacy of users. As a result, we examined privacy issues in terms of data confidentiality and integrity in this section. Only a few mechanisms have been created to meet the need for data privacy. A sender, for example, can obtain privacy by encrypting his data with public key encryption (PKE). Encryption is a means of transforming a communication into another message that can only be decrypted by authorised users. The following are some key approaches that assist data owners in maintaining user privacy while data is kept in the cloud.

6.2 Identity-Based Encryption

As the amount of big data stored in the cloud rises, it becomes more difficult for businesses to maintain data security and confidentiality. On the last two decades, a number of scientists have created a number of identity-based encryption algorithms [18] to protect data privacy in the cloud. Shamir invented Identity-based public-key cryptography in 1984 [19], which generates a public key for encryption and decryption using an individual's IP address, phone number, identity number, or email address. On the other hand, a private key generated by a third-party, referred to as a private key generator (PKG), uses a cryptographic algorithm to compute the related secret / private key from the public key. This strategy is used to protect the user's privacy. When compared to other public-key cryptography systems, an encryption key in IBE is based purely on an individual's identification, which reduces the encryption process's complexity. Despite the fact that some existing technologies can be utilised to update the ciphertext recipient, the Identity-based encryption (IBE) method has a fundamental flaw in that it does not allow ciphertext receiver updation. Consider the following scenario: if the data owner has to make changes to the recipient data, he must first download all data associated with specific users from the cloud, decrypt and re-encrypt the new data, and then outsource the re-encrypted data to the cloud. If the data owner is dealing with large amounts of data, this decryption and re-encryption method can be very time consuming and costly in terms of compute overhead. The data owner must be busy and online at all times throughout the process. As a result, it is necessary to relieve the burden of unnecessary computation on the data owner, which can be accomplished by transferring all computation tasks to a trustworthy third party with the data owner's decryption key. However, this approach causes some anxiety for the data owner because he must be completely reliant on the trustworthy third party, and the security of the encrypted message received by the authenticated user cannot be guaranteed. Mambo and Okamoto [20] proposed the proxy re-encryption (PRE) system, which was further explained in [21]. The PRE mechanism's major purpose is to manage the barriers to information sharing between distinct delivery. Without giving any information about the decryption keys or the actual message, a semi-trusted third party [20] can turn encrypted data for the requested user into encrypted data for other users. This approach transfers the data owner's superfluous workload to the proxy server, which eliminates the need for the proxy server to be online at all times. In [22], a proxy re-encryption method based on identity based encryption (IBE) was created. The authors presented a secret identity-based proxy re-encryption (IBPRE) [23] architecture. The author can only update the receiver ciphertext once using this approach, however the expanding phase of huge data on the cloud necessitates multiple receiver updates. Liang et al. [24] proposed an anonymous identity-based proxy re-encryption strategy to meet the need for numerous ciphertext receiver updates with the likelihood of conditional fine-grained ciphertext sharing while also concealing the sender and receiver's identities.

6.3 Attribute Based Encryption

ABE is now a standard cryptographic tool for controlling fine-grained access to shared data. Several academics have proposed several attribute-based encryption techniques ABE [25], [26] to protect data privacy in the cloud. This approach secures the delivery of selected data in the cloud storage environment from beginning to end.

On ABE, data owners utilise access policies or an access tree structure to establish an environment where only selected data from big data sets is shared with authorised users in the cloud, and that data is encrypted and stored on the cloud under those policies. Encryption of data has been done based on a set of associated attributes to each user. Only those users whose attributes match the structure of the collection of attributes supplied by the data owner can decrypt the data. However, due to the addition or deletion of attributes for specific users, frequent changes in access controls may be required in big data over the cloud by the data owner. Any changes to access policies made by data owners may need to be shared with many entities. As a result, a powerful framework was needed to handle the data owner's random changes in access policy, however the researchers in [27], [28] failed to create an appropriate attribute-based access framework with policy update management techniques. As a result, the creation of such a framework is a difficult process. Following the data outsourcing to the cloud, the data owner deletes all outsourced data from the system and does not keep any replicas on the local system. If the data owner wants to change the existing policy, he must first download all of the data to his local machine, then re-encrypt it using the new policy and upload it back to the cloud server. This entire procedure has a significant communication overhead, is time-consuming, and has a high computing cost. Yang et al. [29] proposed a secure and verifiable updating framework to address the prior work's policy updating issue. The data owner just sends queries to the cloud to make any changes to the existing policy in this framework, and the cloud server performs computation directly on encrypted data to update the policy. This approach does not require downloading all the data and again encrypt it.

The majority of attribute-based access control uses plain text to build the end user's access policy. As a result, these attribute-based access strategies utterly failed to provide data privacy. Various methods introduced by [30]–[34] to hide only the specific values of each attribute, such as wildcards [30], [31], Hidden Vector Encryption [32], and Inner Product Encryption [33], [34], rather than hiding or anonymizing the entire attribute, such as wildcards [30], [31], Hidden Vector Encryption [32], and Inner Product Encryption [33], [34]. While hiding some specific values of attributes can help to protect the user's privacy, the visibility of the attribute name may reveal sensitive information. As a result, the attackers may obtain certain personal information about the user. Consider the case where patient Bob encrypts her data and grants "Cardiologist Doctor" access to her medical information. As a result, the qualities "cardiologist" and "doctor" may play a role in the access policy to Bob's information. Anyone who sees this data, even if he or she is unable to view the concealed value of the characteristic, can infer that Bob may be suffering from heart difficulties, exposing Bob's privacy. As a result, Yang, K. et al. [35] presented a simple approach to hide the entire attributes instead of their corresponding value in the access policy to prevent privacy leaking from the access policy. When the attributes are hidden, however, it is impossible for both authorised and unauthorised users to determine which characteristics are involved in the access policy, making decryption a difficult challenge. Furthermore, the vast majority of these somewhat disguised policy plans are limited to supporting specific policy structures (e.g., AND-gates on multi-valued attributes).

6.4 Homomorphic Encryption

As data has been obtained, it has been kept in either a public or private cloud. Attackers are unlikely to steal information from a private cloud since it is regarded to be trustworthy. As a result, intruders are flocking to the public cloud since it is so easy to compromise privacy. The most popular features, such as multi-tenancy and virtualization, are available in the cloud. As a result, the cloud provider distributes the same physical resources, such as memory space, to several users, increasing the chances of data theft. One of the greatest ways to protect data is to encrypt it before outsourcing it and keep it in the cloud. In recent years, there has been progress in development to achieve higher privacy of data on cloud by Fully Homomorphic encryptions schemes. FHE is an encryption method that allows for random computations on ciphertext data without decrypting the original message [36]. A number of studies have been conducted in the hopes of establishing homomorphic encryption algorithms [37]-[40]. Homomorphic encryption ensures complete data security, but it costs too much to compute and is difficult to create with current technologies. Table 1 shows a comparative analysis of several encryption algorithms.

Table 1.Comparative study on various encryption methods

Encryption Scheme	Features	Limitation
Identity based encryption	• Accessibility of data control is fully dependent on individual characteristic of an user such as email-id, identity number, etc. • It provide full access over all resources	• Time consuming in large environment • Very difficult to design granular access control • Up-dation in encrypted data for specified receiver is impossible • Prior to process data, data must be download and decrypted
Attribute based encryption	• Access control is depend on set of attributes for specified user • It gives more privacy on data and flexibility of access control	• Require excessive computational overhead to handle several user groups • Prior to process data, data must be download and decrypted • Not able to updating ciphertext receiver
Proxy Re-encryption	• Can be deploy on either Identity based encryption or Attribute based encryption environment • Updating encrypted text for receiver is not possible	• Increase computational overhead • Prior to process data, data must be download and decrypted
Homomorphic encryption	• Allow to do computations on ciphertext • Extremely protected schemes	• It give excess computational overhead

6.5 Storage Path Encryption

Storing all of one's data on a centralised storage system in the cloud is one of the most difficult tasks in terms of data privacy leaking, because the cloud storage owner may either leak one's data to a rival or have storage crashes, resulting in data loss. Cheng et al. [12] proposed a strategy to safeguard huge data on cloud storage. Big data is divided into numerous sequenced segments in the author's system, and each segment is gathered on several storage devices held by different cloud service providers. It is necessary to collect all segments from various data storage locations and then restore them to their original state before handing over to the data owner. The storage of big data in the cloud is classified as either public data, private data, or a combination of both in this approach. Because everyone on the network has unrestricted access to public data, it does not require any additional protection. Hidden data, on the other hand, contains sensitive information and must be kept private from unauthorised users. This information is not available to all people and organisations who aren't interested in it.

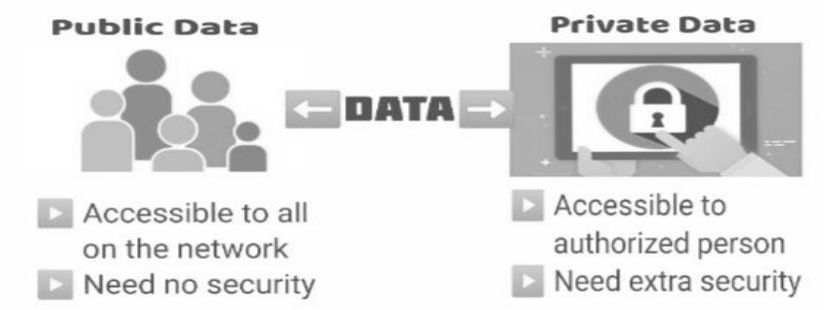

Fig. 5 Difference between Public Data and Private Data

This system includes a trapdoor function that is simple to compute in one direction but difficult to compute in the opposite direction without the use of additional information. This method is mostly employed in cryptography applications. It is preferable to encrypt only the storage path rather than the entire big data, which is known as the cryptographic virtual association of big data. The proposed system also encrypts a few segments of data that are considered private in a few applications. There is a demand to build a method that stores several copies of the same data on separate cloud storage to improve the robustness and availability of large data. It's designed to deal with the unexpected, so that if information or a piece of data is lost from one cloud storage, we can recover it from another cloud. The storage directory information will be maintained by the big data owner [12].

Usage of Hybrid Clouds

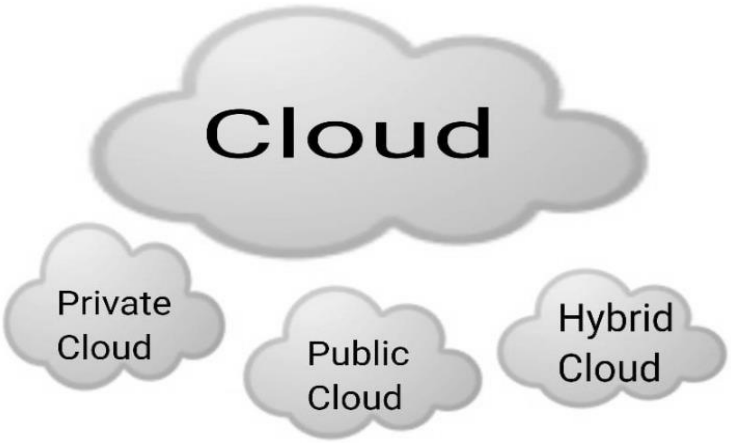

Fig. 6 Standard classificationsof Cloud

The standard definition describes by the National Institute Of Standards And Technology (NIST), the cloud can be categorized into a private cloud, public cloud, and hybrid cloud models [17]. The private clouds which are own by the personal organization and be located behind a firewall. The owner of the private cloud will only decide who will utilize, access, and store data to the cloud from any place. Private

cloud owners give authorization to the user for uses of his private cloud. The public cloud is available to all and can be used by all service subscribers or customers for e.g. Amazon, Microsoft, or Google platform. The hybrid clouds which adopt the features of both public and private clouds. Its ability to combine the scalability of public cloud computing with the security and control of a private cloud. Private clouds are typically secure and trustworthy entities, but there are a few limitations such as scalability, availability, and data sharing that make it difficult to process and store big data [41] on private clouds. More capital investment is required to build such a highly scalable private cloud. In the age of big data, it is difficult to predict the demand for private cloud storage because the size, speed, and variety of data are constantly changing. Another limitation of the private cloud is the lack of software and analytical models required to manage heterogeneous data. In some cases, it may be necessary to share data among authorised users who do not have access to the private cloud or who do not belong to it. However, due to security concerns, sharing is not permitted. As a result, data sharing is yet another limitation of a private cloud. The public cloud, on the other hand, offers greater scalability and easy data sharing but is more vulnerable to security and privacy threats due to multi-tenancy of virtual machines and data. To overcome the limitations of both clouds, hybrid cloud concepts with public cloud and private cloud features have emerged. Researchers have combined the inherent features of the public cloud, such as processing power, scalability, and so on, with the security of the private cloud in this infrastructure. It broadens the scientific research opportunities for big data storage and processing. The majority of the time, hybrid clouds [41] are used for big data storage and privacy-preserving processing. The goal of hybrid clouds is to separate sensitive information from non-sensitive information and collect it in a trusted server, referred to as a private cloud, rather than a trustworthy server, referred to as a public cloud [42].

Integrity Verification of Big Data

Data owners lose physical control over outsourced data when they use cloud computing for big data storage. In a cloud computing environment, the cloud server is assumed to be an untrustworthy entity, putting the outsourced data at risk. As a result, the data owner must be strongly persuaded that his data is properly stored on the cloud in accordance with the service level agreement. To ensure the privacy of cloud users, one solution is to allow them to verify the entire data on the cloud themselves, whether the data is completely stored or not. Another option is to hire a third-party auditor (TPA) to verify the data on behalf of the data owner. As a consequence, there is a need to create a system that achieves an efficient and secure integrity verification mechanism. Several research papers on data integrity verification have been published in the last decade [43-52]. In the last decade, a variety of methods for verifying data integrity, such as the trapdoor hash function, checksum, Reed-Solomon code, message authentication code, and digital signature, have been introduced. One simple method for verifying the integrity of cloud-stored data is to retrieve all of the data from the cloud. However, because of the large volume of big data, it is not efficient in terms of time consumption and communication overhead. To address this issue, data scientists must create an efficient data integrity verification mechanism that does not require downloading whole data from the cloud

[43][44]. When all data is successfully outsourced, the cloud server provides a valid proof of data integrity in integrity verification mechanisms. Integrity verification should be performed on a regular basis to achieve the highest level of protection [43]. The standard architecture of any verification integrity mechanism is depicted in Fig.7, which consists primarily of three entities: Data owner (DO), Cloud Storage Server (CSS), and Third Party Auditor (TPA).

In this architecture, the data owner outsources his data to the cloud, the cloud storage server is responsible for storing the outsourced data, and a third party auditor performs data verification on behalf of the data owner using any data integrity schemes. The integrity of the CSS can be verified by either the data owner or a third-party auditor.
The following steps are taken in the development of a remote integrity verification scheme that supports dynamic data update:
1) Setup and data upload
2) Authorization for TPA
3) Challenge for integrity proof
4) Proof integration
5) Proof verification
6) Updated data upload
7) Updated metadata upload
8) Verification of updated data

Figure 7 depicts the relationship and order of these steps. These steps are explained in detail in [88] as to how they work and why they are necessary for cloud data storage integrity verification.

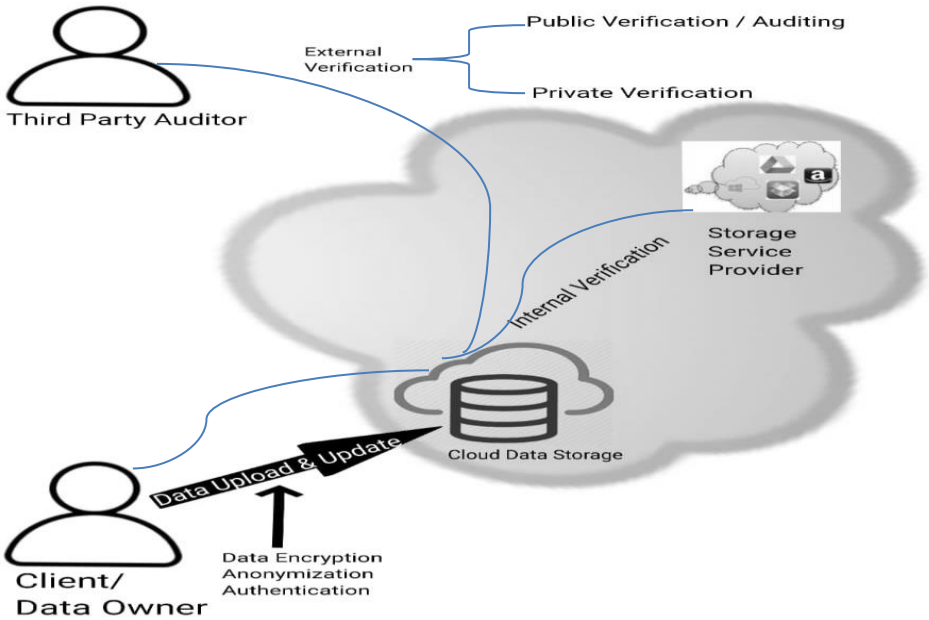

Fig.7 Integrity verification framework for outsourced data

8.1 Provable DataPocession (PDP)

Ateniese et al. [46] proposed the first PDP scheme that allows verification over outsourced data on cloud data storage [48,53]. This scheme also includes blockless verification and public verifiability, both of which can be performed concurrently. The homomorphic verifiable tag (HVT) is used as the building block of PDP based on RSA signature. The HVT tag is stored in the cloud alongside the original file and serves as file block verification metadata. The authors improved the efficiency of the formal S-PDP and dubbed the scheme the PDP light version (E-PDP). However, it was later demonstrated that E-PDP does not provide security under the compact POR model. Integrity verification is a critical area for researchers, so much work must be done to ensure that it can continue to be used by subsequent work such as mixing in random coefficients and probability analysis.

8.2 Proof of Retrievability (POR)

Juels and Kaliski [47] begin by introducing the concept of POR. POR operates on the basis of a test protocol, in which the user receives a response from the service provider to ensure that the file is complete and retrievable. This scheme only works with static data storage, such as a library or an archive. It is also a cryptographic proof scheme that allows a cloud provider to confirm that a user can recover a sought-after file in its entirety. In 2008, Shacham et al. [50] added an improvement to the PORscheme known as compact POR, in which the authors provide enhanced security proof compared to the original PORproof and support the PDP model. They introduced the concept of private verification, in which only the client is authorised to use the private key to verify the data. No other party in the system is authorized to verify it.

8.3 Dynamic Provable Data Processing (DPDA)

The first integrity verification method, Dynamic PDP, was proposed in 2009 [], and it fully supports dynamic data structures. To verify updates, the integrity verification method employs a self-closed life cycle for the processes and a legal data structure such as a rank-based authenticated skip list. When an update such as inert, delete, or modification occurs, a rank-based authenticated skip list uses a logarithm amount of operations, similar to MHT. Except for public verifiability, this scheme has become essential for all dynamic data support mechanisms.

9. Trade-Off Between Utility and Privacy

Generalization and bucketization are the two major anonymization techniques that provide privacy in data publishing, but their use reduces the utility of the data, i.e. a higher degree of data anonymization implies a higher degree of privacy, which has a direct impact on the usefulness of the data, i.e. less knowledge can be extracted from the data. As a result, it is critical to establish an appropriate relationship between privacy and utility in big data. High data privacy results in less data utility, which indicates information loss. Many researchers proposed various computation

strategies to investigate the relationship between privacy gain and utility gain after anonymizing the data. Least distortion [55], visibility metric [56], regularized mean uniformity class size metric [57], and weighted certainty penalty [58], and information theoretic metrics [59], [60] are examples of measuring the data loss schemes. To regulate the relationship between utility and privacy, such as maximise utility, minimise privacy, and maximise privacy where utilities tend to the accuracy in aggregation function estimations. To obtain an appropriate regulate trade-off between utility and privacy, the privacy-preserving data processing used a greedy methodology. During the anonymization process, use privacy preservation metrics and information loss metrics to generate multiple tables that meet the requirements of a specific privacy model. The greedy algorithm produces a table with the least amount of information loss. Measuring data privacy is a difficult task for the data owner. Consider the following scenario: the data owner generates a large amount of data, and a portion of that data is collected by a third party. The third party employs a number of analytical mechanisms to extract new information from the existing database. The data owner has the right to share only those parts of his or her data with a third party that do not reveal his or her sensitive information, as well as to decide how much and what type of data to share. If data is handed over to a third party, the privacy of the data may be jeopardised. If a different data owner manages or handles an account, it provides identical data to a third party. However, when privacy leakage occurs, a few individuals who care deeply about privacy may suffer more loss than those who care little about privacy.

10. Extracting Knowledge From Data

Most organisations use privacy-preserving data mining methods to recognise specific trends and patterns from existing databases in order to learn valuable data from a large amount of existing database without violating privacy. Because big data can be huge, complex, and dynamically changing, those privacy-preserving techniques are ineffective when applied directly to it. To effectively manage such large amounts of data, privacy-preserving techniques may be combined with one or more sets of techniques. Furthermore, those methods should address the privacy concern. To analyse large and complex data, researchers proposed several analytical tools such as clustering, classification, and association rule-based techniques.

10.1 Privacy Preserving Clustering

Clustering is a common data processing method that is used to analyse unused data. Clustering allows you to divide unlabeled input data into distinct sets [61]. Traditional clustering methodology is not supported for big data processing because it requires data to be in a similar layout and filled into a single processing unit. The researchers have published several results in recent years [62], [63]. Despite this, there have been numerous drawbacks related to privacy concerns and computational complexity. Shirkhorshidi et al. [64] proposed a number of methods for various types of clustering. Dimension reduction and sampling methods, in particular, have been developed for centralised machine clustering, while parallel and map-reduce methods have been developed for distributed machine clustering in order to

overcome computational complexity issues. To improve system efficiency, X. Hu. et al. [66] proposed a cloud computing-based parallel processing method. Similarly, Feldman et al. [68] propose a parallel processing method to make clustering suitable for large data sets. A tree construction is used to create core sets, according to this theory. Feldman et al. reduced the required amount of energy while improving processing time. However, privacy is the top priority in all of the methods [61-68]. Because big data involves a large amount of complex data, privacy preservation in clustering becomes a more difficult problem. S. R. M. Oliveira [69] proposed a method for clustering privacy preservation based on hybrid geometric data transformation. This method alters numerical attributes through scaling, rotations, and translations, which improves privacy while decreasing data utility. These methods, however, are not practical. Oliveira and Zaiane [70] describe a method for a centralised data-based system that employs object similarity-based representation and dimensionality reduction. This method was designed for a centralised data system and did not work in a decentralised data system. W. Xiao-Dan [65] proposes privacy-preserving clustering based on a probability distribution-based model to improve clustering efficiency in a new set of data that is unfamiliar with the existing data sets. A. M. Elmisery presents a distributed local clustering method in [67] to handle complex and distributed large data sets. To achieve privacy protection, the author used secure multi-user computation-based methods known as homomorphic encryption. All of the preceding methods used low order statistics for clustering, which produces poor results when the data becomes complex. As a result, low order statistics are ineffective for complex and large data sets. In [71], Shen and Li proposed clustering techniques based on information-theoretic measures to overcome the failure of low order statistics. In this method, nodes are exchanged with their natives based on some parameter rather than actual data.

10.2 Privacy Preserving Data Classification

Classification is the process of assigning a new data item to a predefined group or class. Classification is based on supervised learning concepts, whereas clustering is based on unsupervised learning concepts. The classification algorithms in the traditional approach were designed to work in centralised environments. Because of the tremendous growth in the size of big data, traditional classification algorithms have been enhanced with new features such as the ability to run in a parallel computing environment. For example, in [72], the author introduced a classification algorithm that can process data in two ways: it can classify the data on its own or it can forward the input data to another classifier. This algorithm is known as classification algorithms. This method is capable of performing computations on large and complex data sets very efficiently. Similarly, Rebentrost et al. [73] proposed a quantum-based support vector machine for big data classification. The researcher improved the computational complexity and reduced the required processing information using this method, but it has the disadvantage of using undeveloped hardware technologies in quantum computing. Such efforts in the development of classification algorithms for big data improve performance but cannot be successful in preserving data privacy. Agrawal et al. [74] created a privacy-preserving classification algorithm for extracting information from data. In this method, the

original data is first distorted by appending random offsets, and then the Bayesian formula is used to obtain the density function of the original data, which is then used to rebuild the decision tree. This method's limitation is that it is only appropriate for centralised data. Another method in [75] introduced a privacy-preserving data mining algorithm that employs random reconstruction methods. [76] describes a privacy-preserving method for dealing with distributed databases. To protect the privacy of the data, this method employs the perturbation matrix. The privacy of the original data can be achieved by using a random operation in the algorithm, but this method is not suitable for use with diverse data. As a result, [76] has proposed a privacy-preserving method that can be run on distributed databases. It improves data privacy with the help of the perturbation matrix. Due to the nature of the algorithm, it is necessary to reconstruct the original data from the distorted data set. This method may reduce the algorithm's accuracy. The author of [77] proposed a method for improving the algorithm's accuracy by using a single attribute data random matrix. This method modifies the data slightly, and the reformation of the original data set is aided by the use of a multi-attribute joint distribution matrix. Zhang and Bi [78] developed privacy-preserving techniques for classification by taking advantage of the multi-attribute join distribution matrix and making minor improvements to accuracy and privacy. Despite this, it cannot be appropriate for large and complex data sets.

10.3 Privacy Preserving Association Rule Mining

Several researchers were initially drawn to the data mining-as-a-service (DAAS) model in cloud computing. In this model, data owners outsource their data to the cloud to reduce storage and computational costs, and their mining tasks are delegated to the cloud service provider. This privacy-protection movement began as a serious concern about the success of data mining. Clustering and classification are data mining techniques used to assemble the input data, whereas association rules are used to find the essential patterns or relationships between the input data. Due to the use of parallel computing and cloud computing, traditional methods were incapable of handling large amounts of diverse data. For many years, researchers have been trying to figure out how to discover the interconnected connections of information on larger data sets. Recently, tree structures, such as the FP-tree [79], have been used to obtain the pattern. Several methods [80]-[82] have been developed to handle large and complex data in an efficient manner using the map-reduce technique. In general, the map-reduce technique is appropriate for a cloud-based association rule finding algorithm. Nonetheless, the researchers' proposed methods [80]-[82] do not protect the privacy of the input data. To protect privacy in association rule mining is a method of preventing confidential information from being mined. Several researchers apply privacy-protection concepts, such as in [74], where privacy is achieved by distorting or altering the original data. Changing the original data so that an approximation of the actual data distribution can be generated using distorted data without revealing the meaning of the original data. As a result, [83] has imposed more stringent conditions in order to improve privacy. [84, 85] used privacy protection techniques on Boolean association rules and distorted the original data in a manner similar to other methods. The cryptographic approach is used in some methods to build

decision trees [86]. One example of privacy-preserving data mining is secure multi-party computation [86]. While these methods can protect privacy and accuracy to some extent, they are not fully capable of managing large and complex data sets.

11.Access Control and Secure

11.1 End To End Communication

Data owners offload all sensitive data to the cloud and employ some access control mechanisms to ensure that data is only accessible to authorised users. Before sending data to the cloud, it was encrypted using various encryption techniques such as IBE, ABE, and PRE to ensure data privacy. If the data owner needs to make changes to the data before uploading it, entire datasets must be retrieved or decrypted from the cloud to perform any updated operations. The main issue with retrieving or decrypting entire datasets from the cloud is that the data owner must perform unnecessary downloading, uploading, and time-consuming tasks. Outsourcing re-encrypted updated data to the cloud adds complexity to the data, making it more difficult to perform any computation or fine-grained analysis. Proxy re-encryption schemes have mitigated this disadvantage to some extent. However, in order to obtain the values from the uploaded data, the data must be shared several times with different organisations. Because different organisations use different cryptographic methods, the data owner must decrypt and re-encrypt to generate the decryption keys using their cryptographic mechanisms. These procedures not only add to the computational overhead, but they also increase data insecurity. To address the aforementioned issues, an encryption technique that allows information to be exchanged among multiple data consumers without requiring decryption and re-encryption is preferable.

11.2 Data Anonymization

To protect the data, the data owner must anonymize the data as well as the user's history, which includes all activities performed by authorised users to access the data from the cloud. Data anonymization is a technique for erasing all of a user's records and personal information, which aids in concealing the user's identity. The main issue with anonymization is that the existence of powerful analytic tools and the massive amount of data renders anonymization mechanisms ineffective. Before implementing the anonymization technique, it is necessary to carefully examine "Is anonymized data susceptible to any threats?" and build a data loss metric based on the study of various threat models. We need to develop an effective anonymization mechanism to deal with the dynamic nature of data, as most anonymization techniques only work on static data. This necessitates the development of new utility and privacy metrics. Furthermore, because data anonymization is a time-consuming process, it necessitates automatic data handling with the massive growth of data.

11.3 Decentralized Storage

Traditionally, generated data from various sources was stored in a centralised storage system on the cloud and regarded as third party authority (TPA). Despite the

fact that it has a number of advantages, the failure of a single point results in the loss of all data. The data owner has lost control of the data after outsourcing it to the cloud and is completely reliant on the TPA. Because TPA is assumed to be a semi-trusted authority on the cloud that can disclose personal data to competitors, a single breach in privacy can have far-reaching consequences. As a result, data privacy becomes a difficult task in the cloud environment. To address these issues, several researchers propose implementing multiple authority systems or decentralised storage to store the in multiple cloud. Decentralized systems include IndieWeb [87] and OwnCloud. Data that is extremely sensitive to the organisation will be stored on a private cloud, while non-sensitive data will be stored on the public cloud. Such systems, however, are difficult to manage and increase the managing cost, key generation time, and uploading/downloading time proportionally. As a result, we must devise an efficient system that requires less computation time and is simple to manage.

11.4 Machine Learning Methods and Data Analytics Frameworks

Many powerful analytical tools have been developed by researchers in the fields of machine learning and data mining. These analytical tools are used to predict some valuable information and facts from a large volume of collected datasets and are given full access to the data. Nowadays, various powerful machine learning methods have been applied on cloud data with high execution power (e.g. cloud computing), which has played an important role in big data analytics. These are widely used to influence big data's analytical power. For example, in the fields of astronomy, medical science, and space satellites, the analytical power of big data is being used on a large scale. It is permissible to allow third-party resources to perform analytical computations on sensitive data. Allowing third-party resources to perform analytical computation on sensitive data may result in user privacy violations. To protect users' privacy, machine learning methods such as clustering, classification, and association rule mining must be organised in a privacy-preserving manner.
In some cases, data held by the organization's data owners (e.g., health care data) do not have enough information to find meaningful facts in the same domain, and finding that data may be costly or difficult due to permissible constraints and concerns about privacy violations. To address such issues, most organisations require the deployment of an effective privacy-preserving distributed analytic framework capable of handling disparate datasets from various sources while maintaining the confidentiality of each dataset. Secure data sharing with fine-grained access multiparty computation methods such as homomorphic encryption technique can be used to solve such problems; however, the main issue with using homomorphic encryption in big data is that the analytics must ensure that the computational complexity is kept to a minimum.
Conclusion and Future Research Issues

A massive amount of data is generated and then processed to extract some valuable data from an existing database using various analytical tools, but due to the massive amount of data, it becomes extremely difficult to analyse, process, and secure the data using traditional privacy-preserving methods. It stimulates research interest among researchers and academicians in the field of data integrity verifications. Existing

integrity verification mechanisms have yielded some promising results, and much research is being conducted in the direction of the development of cloud and big data applications that require new emerging systems to meet new requirements and address current challenges in terms of efficiency, storage, computation and communication, security, and scalability / elasticity. Without developing and implementing data-driven algorithms, it is impossible to predict the next generation of big data and cloud computing applications. In this paper, we conducted a thorough investigation and attempted to describe the benefits and drawbacks of each stage of the big data life cycle in terms of privacy and security when it comes to big data applications. Several methods have been introduced in recent years to safeguard the privacy and security of big data, from data generation to data processing, but their efficiency in preserving privacy is remarkable, opening up several issues and challenges.

REFERENCES

[1] J. Gantz and D. Reinsel, Extracting value from chaos, in Proc. IDC I View, Jun. 2011, pp. 1-12.

[2] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. Zürich, Switzerland: McKinsey Global Inst., Jun. 2011, pp. 1-137.

[3] B. Matturdi, X. Zhou, S. Li, and F. Lin, Big data security and privacy: A review, China Commun., vol. 11, no. 14, pp. 135-145, Apr. 2014.

[4] J.K.Laurila, D.Gatica-Perez, I.Aad, J.Blom, O.Bornet, T.M.T.Do, O.Dousse, J.Eberle, M.Miettinen, The mobile data challenge: Big data for mobile computing research, Workshop on the Nokia Mobile Data Challenge, in: Proceedings of the Conjunction with the 10th International ConferenceonPervasiveComputing,2012, pp. 1–8.

[5]J.J. Berman, Introduction, in: Principles of Big Data, Morgan Kaufmann, Boston,2013,xix–xxvi (pp).

[6] D.E. O'Leary, Artificial intelligence and Big data, IEEE Intell. Syst. 28 (2013) 96–99.

[7] M. Chen, S.Mao, Y.Liu, Big data: a survey, Mob. Netw. Appl. 19 (2) (2014) 1–39.

[8] A. Katal, M. Wazid, and R. H. Goudar, Big data: Issues, challenges, tools and good practices, in Proc. IEEE Int. Conf. Contemp. Comput., Aug. 2013, pp. 404-409.

[9] L. Xu, C. Jiang, J.Wang, J. Yuan, and Y. Ren, Information security in big data: Privacy and data mining, in IEEE Access, vol. 2, pp. 1149-1176, Oct. 2014.

[10] H. Hu, Y. Wen, T.-S. Chua, and X. Li, Toward scalable systems for big data analytics: A technology tutorial, IEEE Access, vol. 2, pp. 652-687, Jul. 2014.

[11] Z. Xiao and Y. Xiao, Security and privacy in cloud computing, IEEE Commun. Surveys Tuts., vol. 15, no. 2, pp. 843-859, May 2013.

[12] C. Hongbing, R. Chunming, H. Kai,W.Weihong, and L. Yanyan, Secure big data storage and sharing scheme for cloud tenants, China Commun., vol. 12, no. 6, pp. 106-115, Jun. 2015.

[13] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, Privacy-preserving multikeyword ranked search over encrypted cloud data, IEEE Trans. Parallel Distrib. Syst.,vol. 25, no. 1, pp. 222-233, Jan. 2014.

[14] O. M. Soundararajan, Y. Jenifer, S. Dhivya, and T. K. P. Rajagopal, Data security and privacy in cloud using RC6 and SHA algorithms, Netw. Commun. Eng., vol. 6, no. 5, pp. 202-205, Jun. 2014.

[15] S. Singla and J. Singh, Cloud data security using authentication and encryption technique, Global J. Comput. Sci. Technol., vol. 13, no. 3, pp. 2232-2235, Jul. 2013.
[16] U. Troppens, R. Erkens, W. Muller-Friedt, R. Wolafka, and N. Haustein, Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, iSCSI, InfiniBand and FCoE. New York, NY, USA: Wiley, 2011.
[17] P. Mell and T. Grance, The NIST definition of cloud computing, Nat. Inst. Standards Technol., 2011.
[18] X. Boyen and B.Waters, Anonymous hierarchical identity-based encryption (without random oracles),in Proc. Adv. Cryptol. (ASIACRYPT), vol. 4117. Aug. 2006, pp. 290-307.
[19] Shamir, A. (1984), Identity based cryptosystems and signature schemes, Advances in Cryptology— CRYPTO'84, Lecture Notes in Computer Science, vol. 196, eds. G.R. Blakley and D. Chaum. SpringerVerlag, Berlin.
[20] M. Mambo and E. Okamoto, Proxy cryptosystems: Delegation of the power to decrypt ciphertexts, Fundam. Electron., Commun. Comput. Sci., vol. E80-A, no. 1, pp. 54-63, 1997.
[21] M. Blaze, G. Bleumer, and M. Strauss, Divertible protocols and atomic proxy cryptography, in Proc. Adv. Cryptol. (ASIACRYPT), 1998, pp. 127-144.
[22] M. Green and G. Ateniese, Identity-based proxy re-encryption, in Proc. Int. Conf. Appl. Cryptogr. Netw. Secur., 2007, vol. 4521. pp. 288-306.
[23] J. Shao, Anonymous ID-based proxy re-encryption,in Proc. Int. Conf. Inf. Secur. Privacy, vol. 7372. Jul. 2012, pp. 364-375.
[24] K. Liang, W. Susilo, and J. K. Liu, Privacy-preserving ciphertext multi-sharing control for big data storage, IEEE Trans. Inf. Forensics Security, vol. 10, no. 8, pp. 1578-1589, Aug. 2015.
[25] V. Goyal, O. Pandey, A. Sahai, and B.Waters, Attribute-based encryption for fine-grained access control of encrypted data, in Proc. ACM Conf. Comput. Commun. Secure., Oct. 2006, pp. 89-98.
[26] J. Bethencourt, A. Sahai, and B. Waters, Ciphertext-policy attributebased encryption, in Proc. IEEE Int. Conf. Secur. Privacy, May 2007, pp. 321-334.
[27] K. Yang, X. Jia, K. Ren, B. Zhang, and R. Xie, DAC-MACS: Effective data access control for multiauthority cloud storage systems, IEEE Trans.Inf. Forensics Security, vol. 8, no. 11, pp. 1790-1801, Nov. 2013.
[28] K. Yang and X. Jia, Expressive, efficient, and revocable data access control for multi-authority cloud storage, IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 7, pp. 1735-1744, Jul. 2014.
[29] K. Yang, X. Jia, and K. Ren, Secure and verifiable policy update outsourcing for big data access control in the cloud,IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 12, pp. 3461-3470, Dec. 2015.
[30] T. Nishide, K.Yoneyama, and K. Ohta,Attribute-based encryption with partially hidden encryptor-specified access structures, in Applied cryptography and network security. Springer, 2008, pp. 111–129.
[31] J. Li, K. Ren, B. Zhu, and Z. Wan, Privacy-aware attribute-based encryption with user accountability, in Information Security. Springer, 2009, pp. 347–362.
[32] D. Boneh and B. Waters, Conjunctive, subset, and range queries on encrypted data, in Theory of cryptography. Springer, 2007, pp. 535– 554.

[33] J. Katz, A. Sahai, and B. Waters, Predicate encryption supporting disjunctions, polynomial equations, and inner products, in Advances in Cryptology–EUROCRYPT'08. Springer, 2008, pp. 146–162.
[34] J. Lai, R. H. Deng, and Y. Li, Fully secure cipertext-policy hiding cpabe, in Information Security Practice and Experience. Springer, 2011, pp. 24–39.
[35] K. Yang, Q. Han, H. Li, K. Zheng, Z. Su and X. Shen, An Efficient and Fine-Grained Big Data Access Control Scheme With Privacy-Preserving Policy, in IEEE Internet of Things Journal, vol. 4, no. 2, pp. 563-571, April 2017, doi: 10.1109/JIOT.2016.2571718.
[36] C. Gentry, A fully homomorphic encryption scheme, Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, 2009.
[37]ZvikaBrakerski and Vinod Vaikuntanathan. Fully Homomorphic Encryption from RingLWE and Security for Key Dependent Messages. In Advances in Cryptology - CRYPTO 2011, volume 6841 of Lecture Notes in Computer Science, pages 505–524. Springer, 2011.
[38] ZvikaBrakerski. Fully Homomorphic Encryption without Modulus Switching from Classical GapSVP. IACR Cryptology ePrint Archive, 2012:78, 2012.
[39] ZvikaBrakerski, Craig Gentry, and Vinod Vaikuntanathan. (Leveled) fully homomorphic encryption without bootstrapping. In Innovations in Theoretical Computer Science 2012, pages 309–325. ACM, 2012.
[40] Craig Gentry, Shai Halevi, and Nigel P. Smart. Homomorphic Evaluation of the AES Circuit. IACR Cryptology ePrint Archive, 2012:99, 2012.
[41] S. Nepal, R. Ranjan, and K.-K. R. Choo, Trustworthy processing of healthcare big data in hybrid clouds, IEEE Trans. Cloud Comput., vol. 2, no. 2, pp. 78_84, Mar./Apr. 2015.
[42] X. Huang and X. Du, Achieving big data privacy via hybrid cloud, in Proc. Int. Conf. INFOCOM, Apr. 2014, pp. 512-517.
[43] C. Liu et al., Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates, IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 9, pp. 2234-2244, Sep. 2014.
[44] C. Liu, R. Ranjan, X. Zhang, C. Yang, D. Georgakopoulos, and J. Chen, Public auditing for big data storage in cloud computing. A survey, in Proc. IEEE Int. Conf. Comput. Sci. Eng., Dec. 2013, pp. 1128-1135.
[45] R. C. Merkle, A digital signature based on a conventional encryption function, in Proc. Adv. Cryptol. (CRYPTO), Jan. 1988, pp. 369-378.
[46] G. Ateniese et al., Provable data possession at untrusted stores, in Proc. Int. Conf. ACM Comput. Commun. Secur., 2007, pp. 598-609.
[47] A. Juels and B. S. Kaliski, Jr., PORs: Proofs of retrievability for large files, in Proc. ACM Conf. Comput. Commun. Secur., Oct. 2007, pp. 584-597.
[48] G. Ateniese et al., Remote data checking using provable data possession, Trans. Inf. Syst. Secur., vol. 14, no. 1, May 2011, Art. no. 12.
[49] F. Armknecht, J.-M. Bohli, G. O. Karame, Z. Liu, and C. A. Reuter, Outsourced proofs of retrievability, in Proc. ACM Conf. Comput. Commun. Secur., Nov. 2014, pp. 831-843.
[50] H. Shacham and B. Waters, Compact proofs of retrievability, in Proc. Adv. Cryptol. (ASIACRYPT), Dec. 2008, pp. 90_107.

[51] Q.Wang, C.Wang, K. Ren,W. Lou, and J. Li, Enabling public auditability and data dynamics for storage security in cloud computing, IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 5, pp. 847-859, May 2011.
[52] C. Wang, Q. Wang, K. Ren, and W. Lou, Privacy-preserving public auditing for data storage security in cloud computing, in Proc. IEEE Int. Conf. INFOCOM, Mar. 2010, pp. 1_9.
[53] G. Ateniese, R.B. Johns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, D. Song, Provable data possession at untrusted stores, in: Proceedings of the 14thACM Conference on Computer and Communications Security, CCS'07, 2007, pp. 598–609.
[54]C. Erway, A. Küpçü, C. Papamanthou, R. Tamassia, Dynamic provable datapossession, in: Proceedings of the 16th ACM Conference on Computer andCommunications Security, CCS'09, Chicago, USA, 2009, pp. 213–222.
[55] L. Sweeney, k-anonymity: A model for protecting privacy, Int. J. Uncer- tainty, Fuzziness Knowl. Based Syst., vol. 10, no. 5, pp. 557-570, 2002.
[56] R. J. Bayardo and R. Agrawal, Data privacy through optimal k-anonymization, in Proc. Int. Conf. data Eng., Apr. 2005, pp. 217-228.
[57] K. LeFevre, D. J. Dewitt, and R. Ramakrishnan, Mondrian multidimensional k-anonymity, in Proc. Int. Conf. data Eng., Apr. 2006, p. 25.
[58] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, Utility-based anonymization for privacy preservation with less information loss, ACM SIGKDD Explorations Newslett., vol. 8, no. 2, pp. 21-30, Dec. 2006.
[59] A. Gionis and T. Tassa, k-anonymization with minimal loss of information, IEEE Trans. Knowl. Data Eng., vol. 21, no. 2, pp. 206_219, Feb. 2009.
[60] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. J. Ray Liu, Privacy or utility in data collection? A contract theoretic approach, IEEE J. Sel. Topics SignalProcess., vol. 9, no. 7, pp. 1256-1269, Oct. 2015.
[61] A. K. Jain, M. N. Murty, and P. J. Flynn, Data clustering: A review, ACM Comput. Surv., vol. 31, no. 3, pp. 264-323, Sep. 1999.
[62] R. Xu and D. Wunsch, Clustering. New York, NY, USA: Wiley, 2009.
[63] A. Fahad et al., A survey of clustering algorithms for big data: Taxonomy and empirical analysis, IEEE Trans. Emerg. Topics Comput., vol. 2, no. ,pp. 267-279, Sep. 2014.
[64] A. S. Shirkhorshidi, S. R. Aghabozorgi, Y. W. Teh, and T. Herawan, Big data clustering: A review, in Proc. Int. Conf. Comput. Sci. Appl., 2014, pp. 707_720.
[65] W. Xiao-Dan, Y. Dian-Min, L. Feng-Li, and C. Chao-Hsien,Distributed model based sampling technique for privacy preserving clustering, in Proc. Int. Conf. Manage. Sci. Eng., Aug. 2007, pp. 192-197.
[66] H. Xu, Z. Li, S. Guo, and K. Chen, CloudVista: Interactive and economical visual cluster analysis for big data in the cloud, in Proc. VLDB Endowment, 2012, pp. 1886-1889.
[67] A. M. Elmisery and H. Fu, Privacy preserving distributed learning clustering of healthcare data using cryptography protocols, in Proc. IEEE 34th Annu. Comput. Softw. Appl. Conf. Workshops, Jul. 2010, pp. 140-145.
[68] D. Feldman, M. Schmidt, and C. Sohler, Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering, in Proc. ACM-SIAM Symp. Discrete Algorithms, 2013, pp. 1434-1453.

[69] S. R. M. Oliveira and O. R. Zaiane, Privacy preserving clustering by data transformation, in Proc. 18th Brazilian Symp. Databases, 2003, pp. 304-318.
[70] S. R. M. Oliveira and O. R. Zaiane, Privacy-preserving clustering by object similarity-based representation and dimensionality reduction transformation, in Proc. ICDM Workshop Privacy Security Aspects Data Mining, 2004, pp. 40-46.
[71] P. Shen and C. Li, Distributed information theoretic clustering,IEEE Trans. Signal Process., vol. 62, no. 13, pp. 3442-3453, Jul. 2014.
[72] C. Tekin and M. van der Schaar, Distributed online Big Data classification using context information, in Proc. Int. Conf. Commun., Control, Comput., oct. 2013, pp. 1435-1442.
[73] P. Rebentrost, M. Mohseni, and S. Lloyd. (2014). Quantum Support Vector Machine for Big Feature and Big Data Classification. [Online]. Available:http://dblp.uni-trier.de/db/journals/corr/corr1307.html#RebentrostML13
[74] R. Agrawal and R. Srikant, Privacy-preserving data mining, in Proc. ACM SIGMOD Conf. Manage. Data, 2000, pp. 439_450.
[75] A. Evfimievski, J. Gehrke, and R. Srikant, Limiting privacy breaches in privacy preserving data mining, in Proc. ACM Symp. Principles Database Syst., 2003, pp. 211_222.
[76] S. Agrawal and J. R. Haritsa, A framework for high-accuracy privacypreserving mining, in Proc. 21st Int. Conf. Data Eng., Apr. 2005, pp. 193-204.
[77] G. Weiping, W. Wei, and Z. Haofeng, Privacy preserving classification mining, J. Comput. Res. Develop., vol. 43, no. 1, pp. 39-45, 2006.
[78] X. Zhang and H. Bi, Research on privacy preserving classification data mining based on random peturbation, in Proc. Int. Conf. Inf. Netw. Autom., 2010, pp. 173-178.
[79] J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 1-12.
[80] M.-Y. Lin, P.-Y. Lee, and S.-C. Hsueh, Apriori-based frequent itemset mining algorithms on MapReduce, in Proc. Int. Conf. Ubiquitous Inf. Manage. Commun., 2012, p. 76.
[81] M. Riondato, J. A. DeBrabant, R. Fonseca, and E. Upfal, PARMA: A parallel randomized algorithm for approximate association rules mining in MapReduce, in Proc. ACM Int. Conf. Inf. Knowl. Manage., 2012,pp. 85_94.
[82] C. K.-S. Leung, R. K. MacKinnon, and F. Jiang, Reducing the search space for big data mining for interesting patterns from uncertain data, in Proc. Int. Conf. Big Data, Jun./Jul. 2014, pp. 315_322.
[83] D. Agrawal and C. C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms, in Proc. 20th ACM SIGACT- SIGMOD-SIGART Symp. Principles Database Syst., 2001, pp. 247_255.
[84] A. Ev_mievski, R. Srikant, R. Agrawal, and J. Gehrke, Privacy preserving mining of associationrules, in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 217_228.
[85] S. J. Rizvi and J. R. Haritsa, Maintaining data privacy in association rule mining, in Proc. 28th Int. Conf. Very Large Databases, 2002, pp. 682_693.
[86] Y. Lindell and B. Pinkas, Privacy preserving data mining, in Proc. Adv. Cryptol., 2000, pp. 36_54.

[87] Z.-H. Zhou, N.-V. Chawla, Y. Jin, and G. J. Williams, Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum], IEEE Comput. Intell. Mag., vol. 9, no. 4, pp. 62_74, Nov. 2014.
[88] Liu, C., Yang, C., Zhang, X. & Chen, J. (2015). External integrity verification for outsourced big data in cloud and IoT: A big picture.. Future Generation Comp. Syst., 49, 58-67.
[89] X. Zhang, C. Liu, S. Nepal, S. Panley, J. Chen, A privacy leakage upperboundconstraint based approach for cost-effective privacy preserving ofintermediate datasets in cloud, IEEE Trans. Parallel Distrib. Syst. 24 (2013)1192–1202
[90] K. Zhang, X. Zhou, Y. Chen, X. Wang, Y. Ruan, Sedic: privacy-awaredata intensive computing on hybrid clouds, in: Proceedings of the 18thACM Conference on Computer and Communications Security, CCS'11, 2011,pp. 515–526.
[91] C. Yang, C. Liu, X. Zhang, S. Nepal, J. Chen, A time efficient approach fordetecting errors in big sensor data on cloud, IEEE Trans. Parallel Distrib. Syst.(2013) (in press).
[37-50] H. Shacham, B. Waters, Compact proofs of retrievability, in: Proceedings of the14th International Conference on the Theory and Application of Cryptologyand Information Security, ASIACRYPT'08, 2008, pp. 90–107.
[92] R. Curtmola, O. Khan, R.C. Burns, G. Ateniese, MR-PDP: multiple-replicaprovable data possession, in: Proceedings of the 28th IEEE InternationalConference on Distributed Computing Systems, ICDCS'08, Beijing, China, 2008,pp. 411–420.
[93] C. Erway, A. Küpçü, C. Papamanthou, R. Tamassia, Dynamic provable datapossession, in: Proceedings of the 16th ACM Conference on Computer andCommunications Security, CCS'09, Chicago, USA, 2009, pp. 213–222.
[38-51] Q. Wang, C. Wang, K. Ren, W. Lou, J. Li, Enabling public auditability and datadynamics for storage security in cloud computing, IEEE Trans. Parallel Distrib.Syst. 22 (2011) 847–859.
[94] C. Liu, J. Chen, L.T. Yang, X. Zhang, C. Yang, R. Ranjan, K. Ramamohanarao,Authorized public auditing of dynamic big data storage on cloud with efficientverifiable fine-grained updates, IEEE Trans. Parallel Distrib. Syst. (TPDS) 25(2014) 2234–2244.
[95] C. Liu, R. Ranjan, C. Yang, X. Zhang, L. Wang, J. Chen, MUR-DPA: top–downlevelled multi-replica merkle hash tree based secure public auditing fordynamic big data storage on cloud, IACR Cryptology ePrint Archive, Report2014/391, 2014.