Journal of Big Data

SURVEY PAPER

Open Access

Big data privacy: a technological perspective and review

Priyank Jain*, Manasi Gyanchandani and Nilay Khare

*Correspondence:
priyankjain1984@gmail.com
Computer Science Department, MANIT, Bhopal, India

Abstract

Big data is a term used for very large data sets that have more varied and complex structure. These characteristics usually correlate with additional difficulties in storing, analyzing and applying further procedures or extracting results. Big data analytics is the term used to describe the process of researching massive amounts of complex data in order to reveal hidden patterns or identify secret correlations. However, there is an obvious contradiction between the security and privacy of big data and the widespread use of big data. This paper focuses on privacy and security concerns in big data, differentiates between privacy and security and privacy requirements in big data. This paper covers uses of privacy by taking existing methods such as HybrEx, k-anonymity, T-closeness and L-diversity and its implementation in business. There have been a number of privacy-preserving mechanisms developed for privacy protection at different stages (for example, data generation, data storage, and data processing) of a big data life cycle. The goal of this paper is to provide a major review of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms. This paper also presents recent techniques of privacy preserving in big data like hiding a needle in a haystack, identity based anonymization, differential privacy, privacy-preserving big data publishing and fast anonymization of big data streams. This paper refer privacy and security aspects healthcare in big data. Comparative study between various recent techniques of big data privacy is also done as well.

Keywords: Big data, Privacy and security, Privacy preserving: k-anonymity: T-closeness, L-diversity, HybrEx, PPDP, FADS

Background

Big data [1, 2] specifically refers to data sets that are so large or complex that traditional data processing applications are not sufficient. It's the large volume of data—both structured and unstructured—that inundates a business on a day-to-day basis. Due to recent technological development, the amount of data generated by internet, social networking sites, sensor networks, healthcare applications, and many other companies, is drastically increasing day by day. All the enormous measure of data produced from various sources in multiple formats with very high speed [3] is referred as big data. The term big data [4, 5] is defined as "a new generation of technologies and architectures, designed to economically separate value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery and analysis". On the premise of this definition, the properties of big data are reflected by 3V's, which are, volume, velocity and variety. Later studies pointed out that the definition of 3Vs is insufficient to explain the big data we face now. Thus, veracity, validity, value, variability, venue, vocabulary, and vagueness were added to make some complement explanation of big data [6]. A common theme of big data is that the data are diverse, i.e., they may contain text, audio, image, or video etc. This differing qualities of data is signified by variety. In order to ensure big data privacy, various mechanisms have been developed in recent years. These mechanisms can be grouped based on the stages of big data life cycle [7] Fig. 1, i.e., data generation, storage, and processing. In data generation phase, for the protection of privacy, access restriction as well as falsifying data techniques are used. The approaches to privacy protection in data storage phase are chiefly based on encryption procedures. Encryption based techniques can be further divided into Identity Based Encryption (IBE), Attribute Based Encryption (ABE) and storage path encryption. In addition, to protect the sensitive information, hybrid clouds are utilized where sensitive data are stored in private cloud. The data processing phase incorporates Privacy Preserving Data Publishing (PPDP) and knowledge extraction from the data. In PPDP, anonymization techniques such as generalization and suppression are utilized to protect the privacy of data. These mechanisms can be further divided into clustering, classification and association rule mining based techniques. While clustering and classification split the input data into various groups, association rule mining based techniques find the useful relationships and trends in the input data [8]. To handle diverse measurements of big data in terms of volume, velocity, and variety, there is need to design efficient and effective frameworks to process expansive measure of data arriving at very high speed from various sources. Big data needs to experience multiple phases during its life cycle.

As of 2012, 2.5 quintillion bytes of data are created daily. The volumes of data are vast, the generation speed of data is fast and the data/information space is global [9]. Lightweight incremental algorithms should be considered that are capable of achieving robustness, high accuracy and minimum pre-processing latency. Like, in case of mining, lightweight feature selection method by using Swarm Search and Accelerated PSO can be used in place of the traditional classification methods [10]. Further ahead, Internet of Things (IoT) would lead to connection of all of the things that people care about in the world due to which much more data would be produced than nowadays [11]. Indeed, IoT is one of the major driving forces for big data analytics [9].

In today's digital world, where lots of information is stored in big data's, the analysis of the databases can provide the opportunities to solve big problems of society like

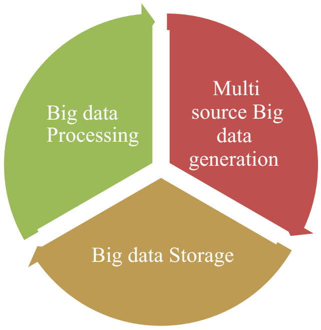

Fig. 1 Big data life cycle stages of big data life cycle, i.e., data generation, storage, and processing are shown

healthcare and others. Smart energy big data analytics is also a very complex and challenging topic that share many common issues with the generic big data analytics. Smart energy big data involve extensively with physical processes where data intelligence can have a huge impact to the safe operation of the systems in real-time [12]. This can also be useful for marketing and other commercial companies to grow their business. As the database contains the personal information, it is vulnerable to provide the direct access to researchers and analysts. Since in this case, the privacy of individuals is leaked, it can cause threat and it is also illegal. The paper is based on research not ranging to a specific timeline. As the references suggest, research papers range from as old as 1998 to papers published in 2016. Also, the number of papers that were retrieved from the keyword-based search can be verified from the presence of references based on the keywords. "Privacy and security concerns" section discusses of privacy and security concerns in big data and "Privacy requirements in big data" section covers the Privacy requirement in big data. "Big data privacy in data generation phase", "Big data privacy in data storage phase" and "Big data privacy preserving in data processing" sections discusses about big data privacy in data generation, data storage, and data processing Phase. "Privacy Preserving Methods in Big Data" section covers the privacy preserving techniques using big data. "Recent Techniques of Privacy Preserving in Big Data" section presents some recent techniques of big data privacy and comparative study between these techniques.

Privacy and security concerns in big data

Privacy and security concerns

Privacy and security in terms of big data is an important issue. Big data security model is not suggested in the event of complex applications due to which it gets disabled by default. However, in its absence, data can always be compromised easily. As such, this section focuses on the privacy and security issues.

Privacy Information privacy is the privilege to have some control over how the personal information is collected and used. Information privacy is the capacity of an individual or group to stop information about themselves from becoming known to people other than those they give the information to. One serious user privacy issue is the identification of personal information during transmission over the Internet [13].

Security Security is the practice of defending information and information assets through the use of technology, processes and training from:-Unauthorized access, Disclosure, Disruption, Modification, Inspection, Recording, and Destruction.

Privacy vs. *security* Data privacy is focused on the use and governance of individual data—things like setting up policies in place to ensure that consumers' personal information is being collected, shared and utilized in appropriate ways. Security concentrates more on protecting data from malicious attacks and the misuse of stolen data for profit [14]. While security is fundamental for protecting data, it's not sufficient for addressing privacy. Table 1 focuses on additional difference between privacy and security.

Privacy requirements in big data

Big data analytics draw in various organizations; a hefty portion of them decide not to utilize these services because of the absence of standard security and privacy protection tools. These sections analyse possible strategies to upgrade big data platforms with the

Table 1 Difference between privacy and security

S.No	Privacy	Security
1	Privacy is the appropriate use of user's information	Security is the "confidentiality, integrity and availability" of data
2	Privacy is the ability to decide what information of an individual goes where	Security offers the ability to be confident that decisions are respected
3	The issue of privacy is one that often applies to a consumer's right to safeguard their information from any other parties	Security may provide for confidentiality. The overall goal of most security system is to protect an enterprise or agency [72]
4	It is possible to have poor privacy and good security practices	However, it is difficult to have good privacy practices without a good data security program
5	For example, if user make a purchase from XYZ Company and provide them payment [13] and address information in order for them to ship the product, they cannot then sell user's information to a third party without prior consent to user	The company XYZ uses various techniques (Encryption, Firewall) in order to prevent data compromise from technology or vulnerabilities in the network

Focuses on additional difference between privacy and security

help of privacy protection capabilities. The foundations and development strategies of a framework that supports:

1. The specification of privacy policies managing the access to data stored into target big data platforms,
2. The generation of productive enforcement monitors for these policies, and
3. The integration of the generated monitors into the target analytics platforms. Enforcement techniques proposed for traditional DBMSs appear inadequate for the big data context due to the strict execution necessities needed to handle large data volumes, the heterogeneity of the data, and the speed at which data must be analysed.

Businesses and government agencies are generating and continuously collecting large amounts of data. The current increased focus on substantial sums of data will undoubtedly create opportunities and avenues to understand the processing of such data over numerous varying domains. But, the potential of big data come with a price; the users' privacy is frequently at danger. Ensures conformance to privacy terms and regulations are constrained in current big data analytics and mining practices. Developers should be able to verify that their applications conform to privacy agreements and that sensitive information is kept private regardless of changes in the applications and/or privacy regulations. To address these challenges, identify a need for new contributions in the areas of formal methods and testing procedures. New paradigms for privacy conformance testing to the four areas of the ETL (Extract, Transform, and Load) process as shown in Fig. 2 [15, 16].

1. *Pre-hadoop process validation* This step does the representation of the data loading process. At this step, the privacy specifications characterize the sensitive pieces of data that can uniquely identify a user or an entity. Privacy terms can likewise indicate which pieces of data can be stored and for how long. At this step, schema restrictions can take place as well.

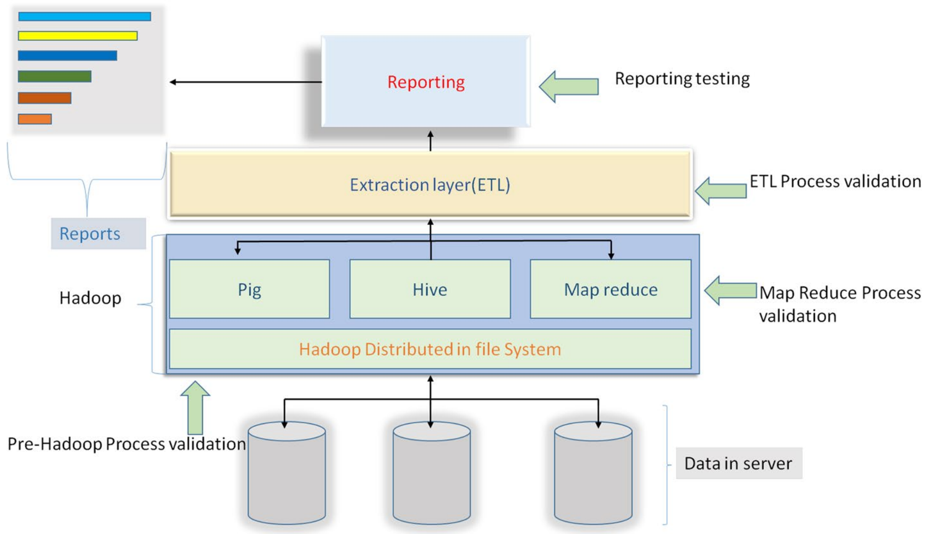

Fig. 2 Big data architecture and testing area new paradigms for privacy conformance testing to the four areas of the ETL (Extract, Transform, and Load) processes are shown here

2. *Map-reduce process validation* This process changes big data assets to effectively react to a query. Privacy terms can tell the minimum number of returned records required to cover individual values, in addition to constraints on data sharing between various processes.
3. *ETL process validation* Similar to step (2), warehousing rationale should be confirmed at this step for compliance with privacy terms. Some data values may be aggregated anonymously or excluded in the warehouse if that indicates high probability of identifying individuals.
4. *Reports testing* reports are another form of questions, conceivably with higher visibility and wider audience. Privacy terms that characterize 'purpose' are fundamental to check that sensitive data is not reported with the exception of specified uses.

Big data privacy in data generation phase

Data generation can be classified into active data generation and passive data generation. By active data generation, we mean that the data owner will give the data to a third party [17], while passive data generation refers to the circumstances that the data are produced by data owner's online actions (e.g., browsing) and the data owner may not know about that the data are being gathered by a third party. Minimization of the risk of privacy violation amid data generation by either restricting the access or by falsifying data.

1. *Access restriction* If the data owner thinks that the data may uncover sensitive information which is not supposed to be shared, it refuse to provide such data. If the data

owner is giving the data passively, a few measures could be taken to ensure privacy, such as anti-tracking extensions, advertisement or script blockers and encryption tools.

2. *Falsifying data* In some circumstances, it is unrealistic to counteract access of sensitive data. In that case, data can be distorted using certain tools prior to the data gotten by some third party. If the data are distorted, the true information cannot be easily revealed. The following techniques are utilized by the data owner to falsify the data:

- A tool Socketpuppet is utilized to hide online identity of individual by deception. By utilizing multiple Socketpuppets, the data belonging to one specific individual will be regarded as having a place with various people. In that way the data collector will not have enough knowledge to relate different socketpuppets to one individual.
- Certain security tools can be used to mask individual's identity, such as Mask Me. This is especially useful when the data owner needs to give the credit card details amid online shopping.

Big data privacy in data storage phase

Storing high volume data is not a major challenge due to the advancement in data storage technologies, for example, the boom in cloud computing [18]. If the big data storage system is compromised, it can be exceptionally destructive as individuals' personal information can be disclosed [19]. In distributed environment, an application may need several datasets from various data centres and therefore confront the challenge of privacy protection.

The conventional security mechanisms to protect data can be divided into four categories. They are file level data security schemes, database level data security schemes, media level security schemes and application level encryption schemes [20]. Responding to the 3V's nature of the big data analytics, the storage infrastructure ought to be scalable. It should have the ability to be configured dynamically to accommodate various applications. One promising technology to address these requirements is storage virtualization, empowered by the emerging cloud computing paradigm [21]. Storage virtualization is process in which numerous network storage devices are combined into what gives off an impression of being a single storage device. SecCloud is one of the models for data security in the cloud that jointly considers both of data storage security and computation auditing security in the cloud [22]. Therefore, there is a limited discussion in case of privacy of data when stored on cloud.

Approaches to privacy preservation storage on cloud

When data are stored on cloud, data security predominantly has three dimensions, confidentiality, integrity and availability [23]. The first two are directly related to privacy of the data i.e., if data confidentiality or integrity is breached it will have a direct effect on users privacy. Availability of information refers to ensuring that authorized parties are able to access the information when needed. A basic requirement for big data storage system is to protect the privacy of an individual. There are some existing mechanisms to fulfil that requirement. For example, a sender can encrypt his data using pubic key encryption (PKE) in a manner that only the valid recipient can decrypt the data. The approaches to safeguard the privacy of the user when data are stored on the cloud are as follows [7]:

- *Attribute based encryption* Access control is based on the identity of a user complete access over all resources.
- *Homomorphic encryption* Can be deployed in IBE or ABE scheme settings updating cipher text receiver is possible.
- *Storage path encryption* It secures storage of big data on clouds.
- *Usage of Hybrid clouds* Hybrid cloud is a cloud computing environment which utilizes a blend of on-premises, private cloud and third-party, public cloud services with organization between the two platforms.

Integrity verification of big data storage

At the point when cloud computing is used for big data storage, data owner loses control over data. The outsourced data are at risk as cloud server may not be completely trusted. The data owner should be firmly convinced that the cloud is storing data properly according to the service level contract. To ensure privacy to the cloud user is to provide the system with the mechanism to allow data owner verify that his data stored on the cloud is intact [24, 25]. The integrity of data storage in traditional systems can be verified through number of ways i.e., Reed-Solomon code, checksums, trapdoor hash functions, message authentication code (MAC), and digital signatures etc. Therefore data integrity verification is of critical importance. It compares different integrity verification schemes discussed [24, 26]. To verify the integrity of the data stored on cloud, straight forward approach is to retrieve all the data from the cloud. To verify the integrity of data without having to retrieve the data from cloud [25, 26]. In integrity verification scheme, the cloud server can only provide the substantial evidence of integrity of data when all the data are intact. It is highly prescribed that the integrity verification should be conducted regularly to provide highest level of data protection [26].

Big data privacy preserving in data processing

Big data processing paradigm categorizes systems into batch, stream, graph, and machine learning processing [27, 28]. For privacy protection in data processing part, division can be done into two phases. In the first phase, the goal is to safeguard information from unsolicited disclosure since the collected data might contain sensitive information of the data owner. In the second phase, the aim is to extract meaningful information from the data without violating the privacy.

Privacy preserving methods in big data

Few traditional methods for privacy preserving in big data is described in brief here. These methods being used traditionally provide privacy to a certain amount but their demerits led to the advent of newer methods.

De-identification

De-identification [29, 30] is a traditional technique for privacy-preserving data mining, where in order to protect individual privacy, data should be first sanitized with generalization (replacing quasi-identifiers with less particular but semantically consistent values) and suppression (not releasing some values at all) before the release for data mining. Mitigate the threats from re-identification; the concepts of k-anonymity [29, 31, 32], l-diversity [30, 31, 33] and t-closeness [29, 33] have been introduced to enhance traditional privacy-preserving data mining. De-identification is a crucial tool in privacy protection, and can be migrated to privacy preserving big data analytics. Nonetheless, as an attacker can possibly get more external information assistance for de-identification in the big data, we have to be aware that big data can also increase the risk of re-identification. As a result, de-identification is not sufficient for protecting big data privacy.

- Privacy-preserving big data analytics is still challenging due to either the issues of flexibility along with effectiveness or the de-identification risks.
- De-identification is more feasible for privacy-preserving big data analytics if develop efficient privacy-preserving algorithms to help mitigate the risk of re-identification.

There are three -privacy-preserving methods of De-identification, namely, K-anonymity, L-diversity and T-closeness. There are some common terms used in the privacy field of these methods:

- *Identifier attributes* include information that uniquely and directly distinguish individuals such as full name, driver license, social security number.
- *Quasi-identifier attributes* means a set of information, for example, gender, age, date of birth, zip code. That can be combined with other external data in order to re-identify individuals.
- *Sensitive attributes* are private and personal information. Examples include, sickness, salary, etc.
- *Insensitive attributes* are the general and the innocuous information.
- *Equivalence classes* are sets of all records that consists of the same values on the quasi-identifiers.

K-anonymity

A release of data is said to have the k-anonymity [29, 31] property if the information for each person contained in the release cannot be perceived from at least k-1 individuals whose information show up in the release. In the context of k-anonymization problems, a database is a table which consists of n rows and m columns, where each row of the table represents a record relating to a particular individual from a populace and the entries in the different rows need not be unique. The values in the different columns are the values of attributes connected with the members of the population. Table 2 is a non-anonymized database comprising of the patient records of some fictitious hospital in Hyderabad.

There are six attributes along with ten records in this data. There are two regular techniques for accomplishing k-anonymity for some value of k.

Table 2 A Non-anonymized database consisting of the patient records

Name	Age	Gender	State of domicile	Religion	Disease
Ramya	29	Female	Tamil Nadu	Hindu	Cancer
Yamini	24	Female	Andhra Pradesh	Hindu	Viral infection
Salini	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joshna	24	Female	Andhra Pradesh	Christian	Heart-related
Badri	23	Male	Karnataka	Buddhist	TB
Ramu	19	Male	Andhra Pradesh	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
John	17	Male	Andhra Pradesh	Christian	Heart-related
Jhonny	19	Male	Andhra Pradesh	Christian	Viral infection

It is a non-anonymized database comprising of the patient records of some fictitious hospital in Hyderabad

1. *Suppression* In this method, certain values of the attributes are supplanted by an asterisk ‘*’. All or some of the values of a column may be replaced by ‘*’. In the anonymized Table 3, replaced all the values in the ‘Name’ attribute and each of the values in the ‘Religion’ attribute by a ‘*’.
2. *Generalization* In this method, individual values of attributes are replaced with a broader category. For instance, the value ‘19’ of the attribute ‘Age’ may be supplanted by ‘ ≤ 20’, the value ‘23’ by ‘$20 < age \leq 30$’, etc.

Table 3 has 2-anonymity with respect to the attributes ‘Age’, ‘Gender’ and ‘State of domicile’ since for any blend of these attributes found in any row of the table there are always no less than two rows with those exact attributes. The attributes that are available to an adversary are called “quasi-identifiers”. Each “quasi-identifier” tuple occurs in at least k records for a dataset with k-anonymity. K-anonymous data can still be helpless against attacks like unsorted matching attack, temporal attack, and complementary release attack [33, 34]. On the positive side, it will present a greedy O(k log k)-approximation algorithm for optimal k-anonymity via suppression of entries. The complexity of rendering relations of private records k-anonymous, while minimizing the amount of information that is not released and simultaneously ensure the anonymity of individuals

Table 3 2-anonymity with respect to the attributes ‘Age’, ‘Gender’ and ‘State of domicile’

Name	Age	Gender	State of domicile	Religion	Disease
*	$20 < Age \leq 30$	Female	Tamil Nadu	*	Cancer
*	$20 < Age \leq 30$	Female	Andhra Pradesh	*	Viral infection
*	$20 < Age \leq 30$	Female	Tamil Nadu	*	TB
*	$20 < Age \leq 30$	Male	Karnataka	*	No illness
*	$20 < Age \leq 30$	Female	Andhra Pradesh	*	Heart-related
*	$20 < Age \leq 30$	Male	Karnataka	*	TB
*	$Age \leq 20$	Male	Andhra Pradesh	*	Cancer
*	$20 < Age \leq 30$	Male	Karnataka	*	Heart-related
*	$Age \leq 20$	Male	Andhra Pradesh	*	Heart-related
*	$Age \leq 20$	Male	Andhra Pradesh	*	Viral infection

Table 3 has 2-Anonymity with respect to the attributes ‘Age’, ‘Gender’ and ‘State of domicile’ since for any blend of these attributes found in any row of the table there are always no less than two rows with those exact attributes up to a group of size k, and withhold a minimum amount of information to achieve this privacy level and this optimization problem is NP-hard. In general, a further restriction of the problem where attributes are suppressed instead of individual entries is also NP-hard [35]. Therefore we move towards L-diversity strategy of data anonymization.

L-diversity

It is a form of group based anonymization that is utilized to safeguard privacy in data sets by reducing the granularity of data representation. This decrease is a trade-off that results outcomes in some loss of viability of data management or mining algorithms for gaining some privacy. The l-diversity model (Distinct, Entropy, Recursive) [29, 31, 34] is an extension of the k-anonymity model which diminishes the granularity of data representation utilizing methods including generalization and suppression in a way that any given record maps onto at least k different records in the data. The l-diversity model handles a few of the weaknesses in the k-anonymity model in which protected identities to the level of k-individuals is not equal to protecting the corresponding sensitive values that were generalized or suppressed, particularly when the sensitive values in a group exhibit homogeneity. The l-diversity model includes the promotion of intra-group diversity for sensitive values in the anonymization mechanism. The problem with this method is that it depends upon the range of sensitive attribute. If want to make data L-diverse though sensitive attribute has not as much as different values, fictitious data to be inserted. This fictitious data will improve the security but may result in problems amid analysis. Also L-diversity method is subject to skewness and similarity attack [34] and thus can't prevent attribute disclosure.

T-closeness

It is a further improvement of l-diversity group based anonymization that is used to preserve privacy in data sets by decreasing the granularity of a data representation. This reduction is a trade-off that results in some loss of adequacy of data management or mining algorithms in order to gain some privacy. The t-closeness model(Equal/Hierarchical distance) [29, 33] extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

An equivalence class is said to have t-closeness if the distance between the conveyance of a sensitive attribute in this class and the distribution of the attribute in the whole table is less than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. The main advantage of t-closeness is that it intercepts attribute disclosure. The problem lies in t-closeness is that as size and variety of data increases, the odds of re-identification too increases. The brute-force approach that examines each possible partition of the table to find the optimal solution takes $\mathrm{n}^{\mathrm{O(n)}}\mathrm{m}^{\mathrm{O(1)}} = 2^{\mathrm{O(nlogn)}}\mathrm{m}^{\mathrm{O(1)}}$ time. We first improve this bound to single exponential in n (Note that it cannot be improved to polynomial unless P = NP) [36].

Comparative analysis of de-identification privacy methods

Advanced data analytics can extricate valuable information from big data but at the same time it poses a big risk to the users' privacy [32]. There have been numerous proposed approaches to preserve privacy before, during, and after analytics process on the big data. This paper discusses three privacy methods such as K-anonymity, L-diversity, and T-closeness. As consumer's data continues to grow rapidly and technologies are unremittingly improving, the trade-off between privacy breaching and preserving will turn out to be more intense. Table 4 presents existing De-identification preserving privacy measures and its limitations in big data.

HybrEx

Hybrid execution model [37] is a model for confidentiality and privacy in cloud computing. It executes public clouds only for operations which are safe while integrating an organization's private cloud, i.e., it utilizes public clouds only for non-sensitive data and computation of an organization classified as public, whereas for an organization's sensitive, private, data and computation, the model utilizes their private cloud. It considers data sensitivity before a job's execution. It provides integration with safety.

The four categories in which HybrEx MapReduce enables new kinds of applications that utilize both public and private clouds are as follows-

1. *Map hybrid* The map phase is executed in both the public and the private clouds while the reduce phase is executed in only one of the clouds as shown in Fig. 3a.

Table 4 Existing De-identification preserving privacy measures and its limitations in big data

S.No	Privacy measure	Definitions	Limitations	Computational complexity
1	K-anonymity	It is a framework for constructing and evaluating algorithms and systems that release information such that released information limits what can be revealed about the properties of entities that are to be protected	Homogeneity-attack, background knowledge	O(k logk) [35, 73]
2	L-diversity	An equivalence class is said to have L-diversity if there are at least "well-represented" values for the sensitive attribute. A table is said to have L-diversity if every equivalence class of the table has L-diversity	L-diversity may be difficult and unnecessary to achieve and L-diversity is insufficient to prevent attribute disclosure	$O((n^2)/k)$
3	T-closeness	An equivalence class is said to have T-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness	T-closeness requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of a sensitive attribute in the overall table	$2^{O(n)O(m)}$ [36]

Presents existing De-identification preserving privacy measures and its limitations in big data along with their computational complexities

2. *Vertical partitioning* It is shown in Fig. 3b. Map and reduce tasks are executed in the public cloud using public data as the input, shuffle intermediate data amongst them, and store the result in the public cloud. The same work is done in the private cloud with private data. The jobs are processed in isolation.
3. *Horizontal partitioning* The Map phase is executed at public clouds only while the reduce phase is executed at a private cloud as can be seen in Fig. 3c.
4. *Hybrid* As in the figure shown in Fig. 3d, the map phase and the reduce phase are executed on both public and private clouds. Data transmission among the clouds is also possible.

Integrity check models of full integrity and quick integrity checking are suggested as well. The problem with HybridEx is that it does not deal with the key that is generated at public and private clouds in the map phase and that it deals with only cloud as an adversary.

Privacy-preserving aggregation

Privacy-preserving aggregation [38] is built on homomorphic encryption used as a popular data collecting technique for event statistics. Given a homomorphic public key encryption algorithm, different sources can use the same public key to encrypt their individual data into cipher texts [39]. These cipher texts can be aggregated, and the aggregated result can be recovered with the corresponding private key. But, aggregation is purpose-specific. So, privacy- preserving aggregation can protect individual privacy in the phases of big data collecting and storing. Because of its inflexibility, it cannot run

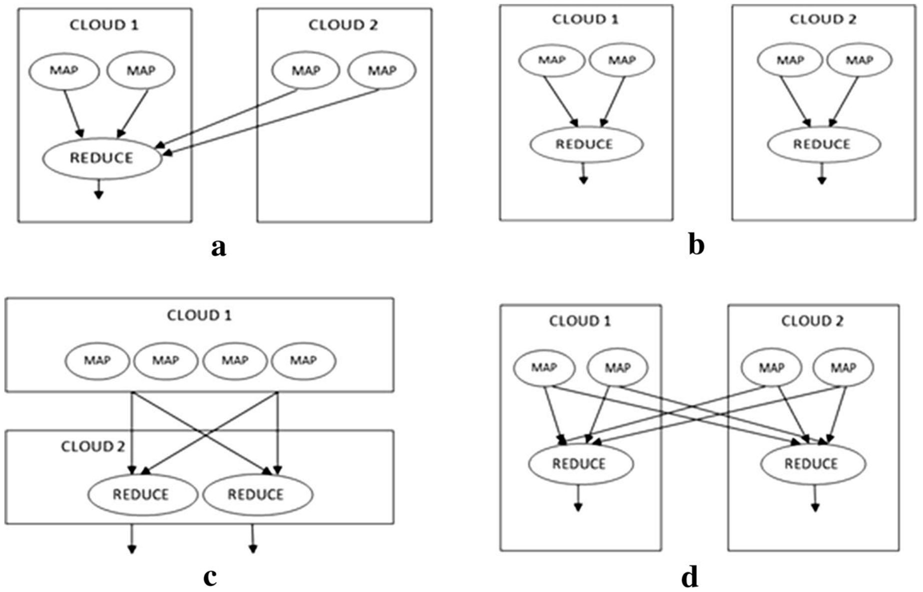

Fig. 3 HybrEx methods: **a** map hybrid **b** vertical partitioning **c** horizontal partitioning **d** hybrid. The four categories in which HybrEx MapReduce enables new kinds of applications that utilize both public and private clouds are shown

complex data mining to exploit new knowledge. As such, privacy-preserving aggregation is insufficient for big data analytics.

Operations over encrypted data

Motivated by searching over encrypted data [38], operations can be run over encrypted data to protect individual privacy in big data analytics. Since, operations over encrypted data are mostly complex along with being time-consuming and big data is high-volume and needs us to mine new knowledge in a reasonable timeframe, running operations over encrypted data can be termed as inefficient in the case of big data analytics.

Recent techniques of privacy preserving in big data

Differential privacy

Differential Privacy [40] is a technology that provides researchers and database analysts a facility to obtain the useful information from the databases that contain personal information of people without revealing the personal identities of the individuals. This is done by introducing a minimum distraction in the information provided by the database system. The distraction introduced is large enough so that they protect the privacy and at the same time small enough so that the information provided to analyst is still useful. Earlier some techniques have been used to protect the privacy, but proved to be unsuccessful.

In mid-90s when the Commonwealth of Massachusetts Group Insurance Commission (GIC) released the anonymous health record of its clients for research to benefit the society [32]. GIC hides some information like name, street address etc. so as to protect their privacy. Latanya Sweeney (then a PhD student in MIT) using the publicly available voter database and database released by GIC, successfully identified the health record by just comparing and co-relating them. Thus hiding some information cannot assures the protection of individual identity.

Differential Privacy (DP) deals to provide the solution to this problem as shown Fig. 4. In DP analyst are not provided the direct access to the database containing personal information. An intermediary piece of software is introduced between the database and the analyst to protect the privacy. This intermediary software is also called as the privacy guard.

Step 1 The analyst can make a query to the database through this intermediary privacy guard.

Step 2 The privacy guard takes the query from the analyst and evaluates this query and other earlier queries for the privacy risk. After evaluation of privacy risk.

Step 3 The privacy guard then gets the answer from the database.

Step 4 Add some distortion to it according to the evaluated privacy risk and finally provide it to the analyst.

The amount of distortion added to the pure data is proportional to the evaluated privacy risk. If the privacy risk is low, distortion added is small enough so that it do not affect the quality of answer, but large enough that they protect the individual privacy of database. But if the privacy risk is high then more distortion is added.

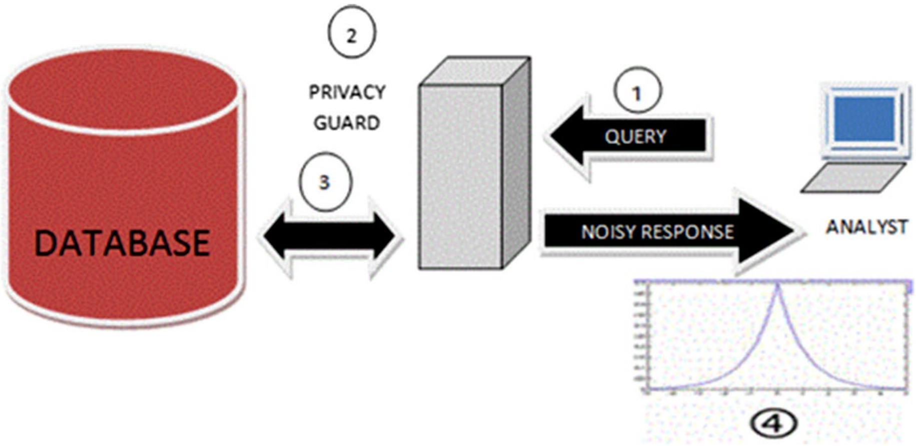

Fig. 4 Differential privacy big data differential privacy (DP) as a solution to privacy-preserving in big data is shown

Identity based anonymization

These techniques encountered issues when successfully combined anonymization, privacy protection, and big data techniques [41] to analyse usage data while protecting the identities of users. Intel Human Factors Engineering team wanted to use web page access logs and big data tools to enhance convenience of Intel's heavily used internal web portal. To protect Intel employees' privacy, they were required to remove personally identifying information (PII) from the portal's usage log repository but in a way that did not influence the utilization of big data tools to do analysis or the ability to re-identify a log entry in order to investigate unusual behaviour. Cloud computing is a type of large-scale distributed computing paradigms which has become a driving force for Information and Communications Technology over the past several years, due to its innovative and promising vision. It provides the possibility of improving IT systems management and is changing the way in which hardware and software are designed, purchased, and utilized. Cloud storage service brings significant benefits to data owners, say, (1) reducing cloud users' burden of storage management and equipment maintenance, (2) avoiding investing a large amount of hardware and software, (3) enabling the data access independent of geographical position, (4) accessing data at any time and from anywhere [42].

To meet these objectives, Intel created an open architecture for anonymization [41] that allowed a variety of tools to be utilized for both de-identifying and re-identifying web log records. In the process of implementing architecture, found that enterprise data has properties different from the standard examples in anonymization literature [43]. This concept showed that big data techniques could yield benefits in the enterprise environment even when working on anonymized data. Intel also found that despite masking obvious Personal Identification Information like usernames and IP addresses, the anonymized data was defenceless against correlation attacks. They explored the trade-offs of correcting these vulnerabilities and found that User Agent (Browser/OS) information strongly correlates to individual users. This is a case study of anonymization implementation in an enterprise, describing requirements, implementation, and experiences encountered when utilizing anonymization to protect privacy in enterprise data analysed using big data techniques. This investigation of the quality of anonymization used k-anonymity based metrics. Intel used Hadoop to analyse the anonymized data and acquire valuable results for the Human Factors analysts [44, 45]. At the same time, learned that anonymization needs to be more than simply masking or generalizing certain fields—anonymized datasets need to be carefully analysed to determine whether they are vulnerable to attack.

Privacy preserving Apriori algorithm in MapReduce framework

Hiding a needle in a haystack [46]

Existing privacy-preserving association rule algorithms modify original transaction data through the noise addition. However, this work maintained the original transaction in the noised transaction in light of the fact that the goal is to prevent data utility deterioration while prevention the privacy violation. Therefore, the possibility that an untrusted cloud service provider infers the real frequent item set remains in the method [47]. Despite the risk of association rule leakage, provide enough privacy protection because this privacy-preserving algorithm is based on "hiding a needle in a haystack" [46] concept. This concept is based on the idea that detecting a rare class of data, such as the needles, is hard to find in a haystack, such as a large size of data, as shown in Fig. 5. Existing techniques [48] cannot add noise haphazardly because of the need to consider privacy-data utility trade-off. Instead, this technique incurs additional computation cost in adding noise that will make the "haystack" to hide the "needle." Therefore, ought to consider a trade-off between problems would be easier to resolve with the use of the Hadoop framework in a cloud environment. In Fig. 5, the dark diamond dots are original association rule and the empty circles are noised association rule. Original rules are hard to be revealed because there are too many noised association rules [46]

In Fig. 6, the service provider adds a dummy item as noise to the original transaction data collected by the data provider. Subsequently, a unique code is assigned to the dummy and the original items. The service provider maintains the code information to filter out the dummy item after the extraction of frequent item set by an external cloud platform. Apriori algorithm is performed by the external cloud platform using data which is sent by the service provider. The external cloud platform returns the frequent item set and support value to the service provider. The service provider filters the frequent item set that is affected by the dummy item using a code to extract the correct association rule using frequent item set without the dummy item. The process of extraction association rule is not a burden to the service provider, considering that the amount of calculation required for extracting the association rule is not much.

Privacy-preserving big data publishing

The publication and dissemination of raw data are crucial components in commercial, academic, and medical applications with an increasing number of open platforms, such as social networks and mobile devices from which data might be gathered, the volume of such data has also increased over time [49]. Privacy-preserving models broadly fall into two different settings, which are referred to as input and output privacy. In input privacy, the primary concern is publishing anonymized data with models such as k-anonymity and l-diversity. In output privacy, generally interest is in problems such as association rule hiding and query auditing where the output of different data mining algorithms is perturbed or audited in order to preserve privacy. Much of the work in privacy has been focused on the quality of privacy preservation (vulnerability quantification) and the utility of the published data. The solution is to just divide the data into smaller parts (fragments) and anonymize each part independently [50].

Despite the fact that k-anonymity can prevent identity attacks, it fails to protect from attribute disclosure attacks because of the lack of diversity in the sensitive attribute within the equivalence class. The l-diversity model mandates that each equivalence class must have at least l well-represented sensitive values. It is common for large data sets to be processed with distributed platforms such as the MapReduce framework [51, 52] in order to distribute a costly process among multiple nodes and accomplish considerable performance improvement. Therefore, in order to resolve the inefficiency, improvements of privacy models are introduced.

Trust evaluation plays an important role in trust management. It is a technical approach of representing trust for digital processing, in which the factors influencing trust are evaluated based on evidence data to get a continuous or discrete number, referred to as a trust value. It propose two schemes to preserve privacy in trust evaluation. To reduce the communication and computation costs, propose to introduce two servers to realize the privacy preservation and evaluation result sharing among various requestors. Consider a scenario with two independent service parties that do not collude with each other due to their business incentives. One is an authorized proxy (AP) that is responsible for access control and management of aggregated evidence to enhance the privacy of entities being evaluated. The other is an evaluation party (EP) (e.g., offered by a cloud service provider) that processes the data collected from a number of trust evidence providers. The EP processes the collected data in an encrypted form and produces an encrypted trust pre-evaluation result. When a user requests the pre-evaluation result from EP, the EP first checks the user's access eligibility with AP. If the check is positive, the AP re-encrypts the pre evaluation result that can be decrypted by the requester (Scheme 1) or there is an additional step involving the EP that prevents the AP from obtaining the plain pre-evaluation result while still allowing decryption of the pre-evaluation result by the requester (Scheme 2) [53].

Improvement of k-anonymity and l-diversity privacy model

MapReduce-based anonymization For efficient data processing MapReduce framework is proposed. Larger data sets are handled with large and distributed MapReduce like frameworks. The data is split into equal sized chunks which are then fed to separate mapper. The mappers process its chunks and provide pairs as outputs. The pairs having the same key are transferred by the framework to one reducer. The reducer output sets are then used to produce the final result [32, 34].

K-anonymity with MapReduce Since the data is automatically split by the MapReduce framework, the k-anonymization algorithm must be insensitive to data distribution across mappers. Our MapReduce based algorithm is reminiscent of the Mondrian algorithm. For better generality and more importantly, reducing the required iterations, each equivalence class is split into (at most) q equivalence classes in each iteration, rather than only two [50].

MapReduce-based l-diversity The extension of the privacy model from k-anonymity to l-diversity requires the integration of sensitive values into either the output keys or values of the mapper. Thus, pairs which are generated by mappers and combiners need to be appropriately modified. Unlike the mapper in k-anonymity, the mapper in l-diversity, receives both quasi-identifiers and the sensitive attribute as input [50].

Fast anonymization of big data streams

Big data associated with time stamp is called big data stream. Sensor data, call centre records, click streams, and health- care data are examples of big data streams. Quality of service (QoS) parameters such as end-to-end delay, accuracy, and real-time processing are some constraints of big data stream processing. The most pre-requirement of big data stream mining in applications such as health-care is privacy preserving [54]. One of the common approaches to anonymize static data is k-anonymity. This approach is not directly applicable for the big data streams anonymization. The reasons are as follows [55]:

1. Unlike static data, data streams need real-time processing and the existing k-anonymity approaches are NP-hard, as proved.
2. For the existing static k-anonymization algorithms to reduce information loss, data must be repeatedly scanned during the anonymization procedure. The same process is impossible in data streams processing.
3. The scales of data streams that need to be anonymized in some applications are increasing tremendously.

Data streams have become so large that anonymizing them is becoming a challenge for existing anonymization algorithms.

To cope with the first and second aforementioned challenges, FADS algorithm was chosen. This algorithm is the best choice for data stream anonymization. But it has two main drawbacks:

1. The FADS algorithm handles tuples sequentially so is not suitable for big data stream.
2. Some tuples may remain in the system for quite a while and are discharged when a specified threshold comes to an end.

This work provided three contributions. First, utilizing parallelism to expand the effectiveness of FADS algorithm and make it applicable for big data stream anonymization. Second, proposal of a simple proactive heuristic estimated round-time to prevent publishing of a tuple after its expiration. Third, illustrating (through experimental results) that FAST is more efficient and effective over FADS and other existing algorithm while it noticeably diminishes the information loss and cost metric during anonymization process.

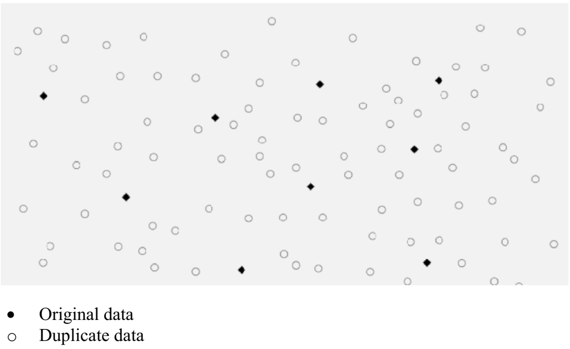

Fig. 5 Hiding a needle in a haystack Mechanism of hiding a needle in a haystack is shown

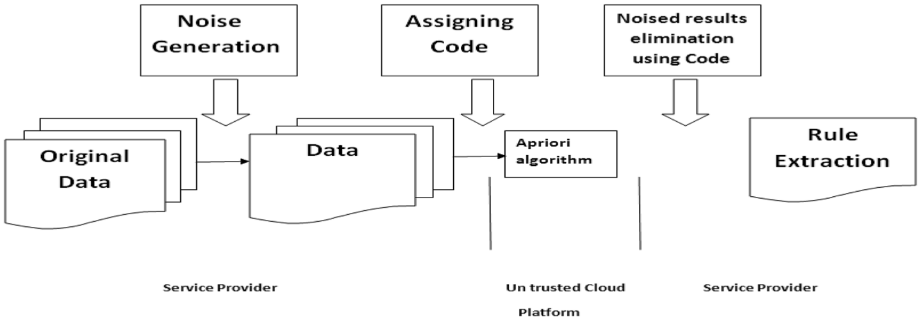

Fig. 6 Overview of the process of association rule mining the service provider adds a dummy item as noise to the original transaction data collected by the data provider

Proactive heuristic

In FADS, a new parameter is considered that represented the maximum delay that is tolerable for an application. This parameter is called expiration-time. To avert a tuple be published when its expiration-time passed, a simple heuristic estimated-round-time is defined. In FADS, there is no check for whether a tuple can remain more in the system or not. As a result, some tuples are published after expiration. This issue is violated the real time condition of a data stream application and also increase cost metric notably.

Privacy and security aspects healthcare in big data

The new wave of digitizing medical records has seen a paradigm shift in the healthcare industry. As a result, healthcare industry is witnessing an increase in sheer volume of data in terms of complexity, diversity and timeliness [56–58]. The term "big data" refers to the agglomeration of large and complex data sets, which exceeds existing computational, storage and communication capabilities of conventional methods or systems. In healthcare, several factors provide the necessary impetus to harness the power of big data [59]. The harnessing the power of big data analysis and genomic research with real-time access to patient records could allow doctors to make informed decisions on treatments [60]. Big data will compel insurers to reassess their predictive models. The real-time remote monitoring of vital signs through embedded sensors (attached to patients) allows health care providers to be alerted in case of an anomaly. Healthcare digitization with integrated analytics is one of the next big waves in healthcare Information Technology (IT) with Electronic Health Records (EHRs) being a crucial building block for this vision. With the introduction of HER incentive programs [61], healthcare organizations recognized EHR's value proposition to facilitate better access to complete, accurate and sharable healthcare data, that eventually lead to improved patient care. With the ever-changing risk environment and introduction of new emerging threats and vulnerabilities, security violations are expected to grow in the coming years [62].

Big data presented a comprehensive survey of different tools and techniques used in Pervasive healthcare in a disease-specific manner. It covered the major diseases and disorders that can be quickly detected and treated with the use of technology, such as fatal and non-fatal falls, Parkinson's disease, cardio-vascular disorders, stress, etc. We have discussed different pervasive healthcare techniques available to address those diseases and many other permanent handicaps, like blindness, motor disabilities, paralysis, etc. Moreover, a plethora of commercially available pervasive healthcare products. It provides understanding of the various aspects of pervasive healthcare with respect to different diseases [63].

Adoption of big data in healthcare significantly increases security and patient privacy concerns. At the outset, patient information is stored in data centres with varying levels of security. Traditional security solutions cannot be directly applied to large and inherently diverse data sets. With the increase in popularity of healthcare cloud solutions, complexity in securing massive distributed Software as a Service (SaaS) solutions increases with varying data sources and formats. Hence, big data governance is necessary prior to exposing data to analytics.

Data governance

1. As the healthcare industry moves towards a value-based business model leveraging healthcare analytics, data governance will be the first step in regulating and managing healthcare data.
2. The goal is to have a common data representation that encompasses industry standards and local and regional standards.
3. Data generated by BSN is diverse in nature and would require normalization, standardization and governance prior to analysis.

Real-time security analytics

1. Analysing security risks and predicting threat sources in real-time is of utmost need in the burgeoning healthcare industry.
2. Healthcare industry is witnessing a deluge of sophisticated attacks ranging from Distributed Denial of Service (DDoS) to stealthy malware.
3. Healthcare industry leverages on emerging big data technologies to make better-informed decisions, security analytics will be at the core of any design for the cloud based SaaS solution hosting protected health information (PHI) [64].

Privacy-preserving analytics

1. Invasion of patient privacy is a growing concern in the domain of big data analytics.

2. Privacy-preserving encryption schemes that allow running prediction algorithms on encrypted data while protecting the identity of a patient is essential for driving healthcare analytics [65].

Data quality

1. Health data is usually collected from different sources with totally different set-ups and database designs which makes the data complex, dirty, with a lot of missing data, and different coding standards for the same fields.
2. Problematic handwritings are no more applicable in EHR systems, the data collected via these systems are not mainly gathered for analytical purposes and contain many issues—missing data, incorrectness, miscoding—due to clinicians' workloads, not user friendly user interfaces, and no validity checks by humans [66].

Data sharing and privacy

1. The health data contains personal health information (PHI), there will be legal difficulties in accessing the data due to the risk of invading the privacy.
2. Health data can be anonymized using masking and de-identification techniques, and be disclosed to the researchers based on a legal data sharing agreement [67].
3. The data gets anonymized so much with the aim of protecting the privacy, on the other hand it will lose its quality and would not be useful for analysis anymore And coming up with a balance between the privacy-protection elements (anonymization, sharing agreement, and security controls) is essential to be able to access a data that is usable for analytics.

Relying on predictive models

1. It should not be unrealistic expectations from the constructed data mining models. Every model has an accuracy.
2. It is important to consider that it would be dangerous to only rely on the predictive models when making critical decisions that directly affects the patient's life, and this should not even be expected from the predictive model.

Variety of methods and complex math's

1. The underlying math of almost all data mining techniques is complex and not very easily understandable for non-technical fellows, thus, clinicians and epidemiologists have usually preferred to continue working with traditional statistics methods.
2. It is essential for the data analyst to be familiar with the different techniques, and also the different accuracy measurements to apply multiple techniques when analysing a specific dataset.

Summary on recent approaches used in big data privacy

In this section, a summary on recent approaches used in big data privacy is done. Table 5 is presented here comprising of different papers, the methods introduced, their focus and demerits. It presents an overview of the work done till now in the field of big data privacy.

Table 5 Summary on recent approaches used in big data privacy

S.No	Research paper	Publication and year	Focus	Limitations
1	"Toward Efficient and Privacy Preserving Computing in Big Data Era" [38]	IEEE Network July/Aug 2014	Introduced an efficient and privacy-preserving cosine similarity computing protocol	Need significant research efforts for addressing unique Privacy issues in some specific big data analytics
2	"Hiding a needle in a Haystack: privacy preserving Apriori algorithm in map reduce framework" [46]	ACM Nov 7, 2014	Proposed the privacy preserving data mining technique in Hadoop i.e. solve privacy violation without utility degradation	Execution time of proposed technique is affected by noise size
3	"Making big data, privacy, and anonymization work together in the enterprise: experiences and issues" [41]	IEEE International Congress 2014	Discusses experiences and issues encountered when successfully combined anonymization, privacy protection, and Big Data techniques to analyse usage data while protecting the identities of users	Uses K-anonymity technique which is vulnerable to correlation attack
4	"Microsoft Differential Privacy for Everyone" [40]	Microsoft Research 2015	Discussed and suggested how an existing approach "differential privacy" is suitable for big data	This method total depends on calculation of the amount of noise by the curator. So if curator is compromised the whole system fails
5	"A scalable two-phase top-down specialization approach for data anonymization using MapReduce on cloud" [69]	IEEE transactions on parallel and distributed systems 2014	Proposed a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the Map Reduce framework on cloud	Uses anonymization technique which is vulnerable to correlation attack
6	"HireSome-II: towards privacy-aware cross-cloud service composition for big data applications" [74]	IEEE transactions on parallel and distributed systems 2014	Proposed a privacy-aware cross-cloud service composition method, named HireSome-II (History record-based Service optimization method) based on its previous basic version HireSome-I	
7	Protection of big data privacy [7]	IEEE translations 2016	Proposed various privacy issues dealing with big data applications	Customer segmentation and profiling can easily lead to discrimination based on age gender, ethnic background, health condition, social, background, and so on
8	Fast anonymization of big data streams [55]	ACM August, 2014	Proposed an anonymization algorithm (FAST) to speed up anonymization of big data streams	Further research required to design and implement FAST in a distributed cloud-based framework in order to gain cloud computation power and achieve high scalability

Table 5 continued

S.No	Research paper	Publication and year	Focus	Limitations
9	Privacy preserving Ciphertext multi-sharing control for big data storage [75]	IEEE Transactions on informatics Forensics and Security 2015	Proposed a privacy-preserving Ciphertext multi-sharing mechanism	The proxy can create delegation rights between the two parties which have never agreed upon the delegation process
10	Privacy-preserving machine learning algorithms for big data systems [76]	IEEE international conference on distributed computing systems 2015	Proposed a novel framework to achieve privacy-preserving machine learning where the training data are distributed and each shared data portion of large volume	Not able to achieve distributed feature selection
11	Privacy-preserving big data publishing [50]	ACM June–July 2015	Proposed approach towards privacy-preserving data mining of very massive data sets using MapReduce	Generalization is unable to handle high dimensional data, it reduces data utility. Perturbation reduces utility of data
12	Proximity-aware local-recoding anonymization with map reduce for scalable big data privacy preservation in cloud [70]	IEEE Transactions on computer August 2015	Model the problem of big data local recoding against proximity privacy breaches as a proximity-aware clustering problem, and propose a scalable two-phase clustering approach accordingly	Further research to integrate our approach with Apache Mahout to achieve highly scalable privacy preserving big data mining or analytics
13	Deduplication on encrypted big data in cloud [77]	IEEE transactions on big data 2016	Proposed a practical scheme to manage the encrypted big data in cloud with deduplication based on ownership challenge and Proxy Re-Encryption (PRE)	Convergent encryption(CE) is subject to an inherent security limitation, namely, susceptibility to offline Brute-force dictionary attacks
14	Security and privacy for storage and computation in cloud computing [22]	International Journal of Science and Research (IJSR) ISSN (Online): 2319–7064	Proposed methodology provides data confidentiality, secure data sharing without Re-encryption, access control for malicious insiders, and forward and backward access control	Limiting the trust level in the cryptographic server (CS)

Provides a list of papers with emphasis on their focus and limitation

Conclusion and future work

Big data [2, 68] is analysed for bits of knowledge that leads to better decisions and strategic moves for overpowering businesses. Yet only a small percentage of data is actually analysed. In this paper, we have investigated the privacy challenges in big data by first identifying big data privacy requirements and then discussing whether existing privacy-preserving techniques are sufficient for big data processing. Privacy challenges in each phase of big data life cycle [7] are presented along with the advantages and disadvantages of existing privacy-preserving technologies in the context of big data applications. This paper also presents traditional as well as recent techniques of privacy preserving in big data. Hiding a needle in a haystack [46] is one such example in which privacy preserving is used by association rule mining. Concepts of identity based anonymization [41] and differential privacy [40] and comparative study between various recent techniques of big data privacy are also discussed. It presents scalable anonymization methods [69] within the MapReduce framework. It can be easily scaled up by increasing the number of mappers and reducers. As our future direction, perspectives are needed to achieve effective solutions to the scalability problem [70] of privacy and security in the era of big data and especially to the problem of reconciling security and privacy models by exploiting the map reduce framework. In terms of healthcare services [59, 64–67] as well, more efficient privacy techniques need to be developed. Differential privacy is one such sphere which has got much of hidden potential to be utilized further. Also with the rapid development of IoT, there are lots of challenges when IoT and big data come; the quantity of data is big but the quality is low and the data are various from different data sources inherently possessing a great many different types and representation forms, and the data is heterogeneous, as-structured, semi structured, and even entirely unstructured [71]. This poses new privacy challenges and open research issues. So, different methods of privacy preserving mining may be studied and implemented in future. As such, there exists a huge scope for further research in privacy preserving methods in big data.

Authors' contributions

PJ performed the primary literature review and analysis for this manuscript work. MG worked with PJ to develop the article framework and focus, and MG also drafted the manuscript. NK introduced this topic to PJ. MG and NK revised the manuscript for important intellectual content and have given final approval of the version to be published. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 27 July 2016 Accepted: 5 November 2016
Published online: 26 November 2016

References

1. Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. Aurora: a new model and architecture for data stream manag ement. VLDB J. 2003;12(2):120–39.
2. Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S. An efficient time optimized scheme for progressive analytics in big data. Big Data Res. 2015;2(4):155–65.
3. Big data at the speed of business, [online]. http://www-01.ibm.com/soft-ware/data/bigdata/2012.
4. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. Big data: the next frontier for innovation, competition, and productivity. New York: Mickensy Global Institute; 2011. p. 1–137.
5. Gantz J, Reinsel D. Extracting value from chaos. In: Proc on IDC IView. 2011. p. 1–12.
6. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data Springer Open J. 2015.
7. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and content mining are permitted for academic research. 2016.

8. Jain P, Pathak N, Tapashetti P, Umesh AS. Privacy preserving processing of data decision tree based on sample selection and singular value decomposition. In: 39th international conference on information assurance and security (IAS). 2013.
9. Qin Y, et al. When things matter: a survey on data-centric internet of things. J Netw Comp Appl. 2016;64:137–53.
10. Fong S, Wong R, Vasilakos AV. Accelerated PSO swarm search feature selection for data stream mining big data. In: IEEE transactions on services computing, vol. 9, no. 1. 2016.
11. Middleton P, Kjeldsen P, Tully J. Forecast: the internet of things, worldwide. Stamford: Gartner; 2013.
12. Hu J, Vasilakos AV. Energy Big data analytics and security: challenges and opportunities. IEEE Trans Smart Grid. 2016;7(5):2423–36.
13. Porambage P, et al. The quest for privacy in the internet of things. IEEE Cloud Comp. 2016;3(2):36–45.
14. Jing Q, et al. Security of the internet of things: perspectives and challenges. Wirel Netw. 2014;20(8):2481–501.
15. Han J, Ishii M, Makino H. A hadoop performance model for multi-rack clusters. In: IEEE 5th international conference on computer science and information technology (CSIT). 2013. p. 265–74.
16. Gudipati M, Rao S, Mohan ND, Gajja NK. Big data: testing approach to overcome quality challenges. Data Eng. 2012:23–31.
17. Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. IEEE Access. 2014;2:1149–76.
18. Liu S. Exploring the future of computing. IT Prof. 2011;15(1):2–3.
19. Sokolova M, Matwin S. Personal privacy protection in time of big data. Berlin: Springer; 2015.
20. Cheng H, Rong C, Hwang K, Wang W, Li Y. Secure big data storage and sharing scheme for cloud tenants. China Commun. 2015;12(6):106–15.
21. Mell P, Grance T. The NIST definition of cloud computing. Natl Inst Stand Technol. 2009;53(6):50.
22. Wei L, Zhu H, Cao Z, Dong X, Jia W, Chen Y, Vasilakos AV. Security and privacy for storage and computation in cloud computing. Inf Sci. 2014;258:371–86.
23. Xiao Z, Xiao Y. Security and privacy in cloud computing. In: IEEE Trans on communications surveys and tutorials, vol 15, no. 2, 2013. p. 843–59.
24. Wang C, Wang Q, Ren K, Lou W. Privacy-preserving public auditing for data storage security in cloud computing. In: Proc. of IEEE Int. Conf. on INFOCOM. 2010. p. 1–9.
25. Liu C, Ranjan R, Zhang X, Yang C, Georgakopoulos D, Chen J. Public auditing for big data storage in cloud computing—a survey. In: Proc. of IEEE Int. Conf. on computational science and engineering. 2013. p. 1128–35.
26. Liu C, Chen J, Yang LT, Zhang X, Yang C, Ranjan R, Rao K. Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates. In: IEEE trans. on parallel and distributed systems, vol 25, no. 9. 2014. p. 2234–44
27. Xu K, et al. Privacy-preserving machine learning algorithms for big data systems. In: Distributed computing systems (ICDCS) IEEE 35th international conference; 2015.
28. Zhang Y, Cao T, Li S, Tian X, Yuan L, Jia H, Vasilakos AV. Parallel processing systems for big data: a survey. In: Proceedings of the IEEE. 2016.
29. Li N, et al. t-Closeness: privacy beyond *k*-anonymity and *L*-diversity. In: Data engineering (ICDE) IEEE 23rd international conference; 2007.
30. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond *k-anonymity.* In: Proc. 22nd international conference data engineering (ICDE); 2006. p. 24.
31. Ton A, Saravanan M. Ericsson research. [Online]. http://www.ericsson.com/research-blog/data-knowledge/big-data-privacy-preservation/2015.
32. Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.
33. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory; 1998.
34. Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertain Fuzz. 2002;10(5):557–70.
35. Meyerson A, Williams R. On the complexity of optimal k-anonymity. In: Proc. of the ACM Symp. on principles of database systems. 2004.
36. Bredereck R, Nichterlein A, Niedermeier R, Philip G. The effect of homogeneity on the complexity of k-anonymity. In: FCT; 2011. p. 53–64.
37. Ko SY, Jeon K, Morales R. The HybrEx model for confidentiality and privacy in cloud computing. In: 3rd USENIX workshop on hot topics in cloud computing, HotCloud'11, Portland; 2011.
38. Lu R, Zhu H, Liu X, Liu JK, Shao J. Toward efficient and privacy-preserving computing in big data era. IEEE Netw. 2014;28:46–50.
39. Paillier P. Public-key cryptosystems based on composite degree residuosity classes. In: EUROCRYPT. 1999. p. 223–38.
40. Microsoft differential privacy for everyone, [online]. 2015. http://download.microsoft.com/…/Differential_Privacy_for_Everyone.pdf.
41. Sedayao J, Bhardwaj R. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. Big Data Congress; 2014.
42. Yong Yu, et al. Cloud data integrity checking with an identity-based auditing mechanism from RSA. Future Gener Comp Syst. 2016;62:85–91.
43. Oracle Big Data for the Enterprise, 2012. [online]. http://www.oracle.com/ca-en/technoloqies/biq-doto.
44. Hadoop Tutorials. 2012. https://developer.yahoo.com/hadoop/tutorial.
45. Fair Scheduler Guide. 2013. http://hadoop.apache.org/docs/r0.20.2/fair_scheduler.html,
46. Jung K, Park S, Park S. Hiding a needle in a haystack: privacy preserving Apriori algorithm in MapReduce framework PSBD'14, Shanghai; 2014. p. 11–17.
47. Ateniese G, Johns RB, Curtmola R, Herring J, Kissner L, Peterson Z, Song D. Provable data possession at untrusted stores. In: Proc. of int. conf. of ACM on computer and communications security. 2007. p. 598–609.
48. Verma A, Cherkasova L, Campbell RH. Play it again, SimMR!. In: Proc. IEEE Int'l conf. cluster computing (Cluster'11); 2011.

49. Feng Z, et al. TRAC: Truthful auction for location-aware collaborative sensing in mobile crowd sourcing INFOCOM. Piscataway: IEEE; 2014. p. 1231–39.
50. HessamZakerdah CC, Aggarwal KB. Privacy-preserving big data publishing. La Jolla: ACM; 2015.
51. Dean J, Ghemawat S. Map reduce: simplied data processing on large clusters. OSDI; 2004.
52. Lammel R. Google's MapReduce programming model-revisited. Sci Comput Progr. 2008;70(1):1–30.
53. Yan Z, et al. Two schemes of privacy-preserving trust evaluation. Future Gener Comp Syst. 2016;62:175–89.
54. Zhang Y, Fong S, Fiaidhi S, Mohammed S. Real-time clinical decision support system with data stream mining. J Biomed Biotechnol. 2012;2012:8.
55. Mohammadian E, Noferesti M, Jalili R. FAST: fast anonymization of big data streams. In: ACM proceedings of the 2014 international conference on big data science and computing, article 1. 2014.
56. Haferlach T, Kohlmann A, Wieczorek L, Basso G, Kronnie GT, Bene M-C, De Vos J, Hernandez JM, Hofmann W-K, Mills KI, Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Liu WM, Williams PM, Fo R. Clinical utility of microarray-based gene expression profiling in the diagnosis and sub classification of leukemia: report from the international microarray innovations in leukemia study group. J Clin Oncol. 2010;28(15):2529–37.
57. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Tollenaar R. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol. 2011;29(1):17–24.
58. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–7.
59. Groves P, Kayyali B, Knott D, Kuiken SV. The 'big data' revolution in healthcare. New York: McKinsey & Company; 2013.
60. Public Law 111–148—Patient Protection and Affordable Care Act. U.S. Government Printing Office (GPO); 2013.
61. EHR incentive programs. 2014. [Online]. https://www.cms.gov/Regulations-and Guidance/Legislation/EHRIncentive-Programs/index.html.
62. First things first—highmark makes healthcare-fraud prevention top priority with SAS. SAS; 2006.
63. Acampora G, et al. Data analytics for pervasive health. In: Healthcare data analytics. ISSN:533-576. 2015.
64. Haselton MG, Nettle D, Andrews PW. The evolution of cognitive bias. In: The handbook of evolutionary psychology. Hoboken: Wiley; 2005. p. 724–46.
65. Hill K. How target figured out a teen girl was pregnant before her father did. New York: Forbes, Inc.; 2012. [Online]. http://www.forbes.com/sites/kashmirhill/2012/02/16/howtarget- figured-out-a-teen-girl-was-pregnant-before-herfather- did/.
66. Violán C, Foguet-Boreu Q, Hermosilla-Pérez E, Valderas JM, Bolíbar B, Fàbregas-Escurriola M, Brugulat-Guiteras P, Muñoz-Pérez MÁ. Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multi morbidity. BMC Public Health. 2013;13(1):251.
67. Emam KE. Guide to the de-identification of personal health information. Boca Raton: CRC Press; 2013.
68. Wu X. Data mining with big data. IEEE Trans Knowl Data Eng. 2014;26(1):97–107.
69. Zhang X, Yang T, Liu C, Chen J. A scalable two-phase top-down specialization approach for data anonymization using systems, in MapReduce on cloud. IEEE Trans Parallel Distrib. 2014;25(2):363–73.
70. Zhang X, Dou W, Pei J, Nepal S, Yang C, Liu C, Chen J. Proximity-aware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud. In: IEEE transactions on computers, vol. 64, no. 8, 2015.
71. Chen F, et al. Data mining for the internet of things: literature review and challenges. Int J Distrib Sens Netw. 2015;501:431047.
72. Fei H, et al. Robust cyber-physical systems: concept, models, and implementation. Future Gener Comp Syst. 2016;56:449–75.
73. Sweeney L. k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst. 2002;10(5):557–70.
74. Dou W, et al. Hiresome-II: towards privacy-aware cross-cloud service composition for big data applications. IEEE Trans Parallel Distrib Syst. 2014;26(2):455–66.
75. Liang K, Susilo W, Liu JK. Privacy-preserving ciphertext for big data storage. In: IEEE transactions on informatics and forensics security. vol 10, no. 8. 2015.
76. Xu K, Yue H, Guo Y, Fang Y. Privacy-preserving machine learning algorithms for big data systems. In: IEEE 35th international conference on distributed systems. 2015.
77. Yan Z, Ding W, Xixun Yu, Zhu H, Deng RH. Deduplication on encrypted big data in cloud. IEEE Trans Big Data. 2016;2(2):138–50.