I. J. Computer Network and Information Security, 2022, 5, 11-27
Published Online on October 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2022.05.02

Sensitive Data Identification and Security Assurance in Cloud and IoT based Networks

Soumya Ray
Dept. of CSE, B.I.T. Mesra, India
E-mail: soumyaray@bitmesra.ac.in

Kamta Nath Mishra*
Dept. of CSE, B.I.T. Mesra, India
Corresponding Author E-mail: mishrakn@yahoo.com

Sandip Dutta
Dept. of CSE, B.I.T. Mesra, India
E-mail: sandipdutta@bitmesra.ac.in

Received: 21 June 2021; Revised: 15 December 2021; Accepted: 08 May 2022; Published: 08 October 2022

Abstract: Sensitive data identification is a vital strategy in any distributed system. However, in the case of non-appropriate utilization of the system, sensitive data security can be at risk. Therefore, sensitive data identification and its security validation are mandatory. The paper primarily focuses on novel sensitive data recognition methodologies. Further, the sensitivity score of the attributes distinguishes non-sensitive attributes, and domain expert plays an important role in this process. The designing of the security assurance Algo and their corresponding decision tables make the system more robust and reliable. The result section is validated with the help of graphical representation, which clearly makes the authenticity of the research work. In summary, the authors may say that the sensitive data identification and security assurance of the proposed system is automated and work optimally in a cloud-based system.

Index Terms: Moving Sensitive Data, Distributed Computing, Cloud-IoT Technology, Quasi technique, Static Sensitive Data, Security Surveillance.

1. Introduction

Now a day's, cloud-computing system is inevitable to maintain the storage of large data on request from different sources. The cloud service provider provides service to the customers on the basis of payment. Therefore, software, platform, and infrastructure can easily be accessed at a lower rate. The major advantage of cloud computing is that sharing of data is possible in a collaborative way, thus ultimately turning the cloud into a globally accepted technology for different industries [1-5]. However, various clients are not interested in taking the service due to its lack of security and trustworthiness. Large organizations store their financial transactions in the cloud, e.g., Goldman Sachs [6]. Most of the data is generated through different heterogeneous IoT devices. These data are very sensitive and secured data. Processing and transferring encrypted data through a net consume huge time and bandwidth. Encrypted data is not fruitful due to its size and low transfer rate. Sensitive data identification is very important to mitigate security issues. Sensitive data identification can be possible using the quasi technique [7-10].

Quasi-identifier is the combination of tuples that uniquely identify individuals by connecting external data. Without an individual's name, the person can be identified by the Voter's ID or SSN, etc. [11]. The identification of a person is possible with a combination of attributes, e.g., age, gender, and location. This is possible when the published data can be joined or merged with data sets listing specific knowledge about these identities. Qusai-Identifier contains information that can easily identify an individual entity when it is joined with other category information. They have not been treated as direct identifiers [12-15]. QIS is helpful in finding the sensitivity attribute as it considers the sensitivity score given by the domain expert. The range of sensitivity scores can also be applied with the quasi-identifier technique to categorize the different sensitive data for further processing. A large amount of data loss is happened due to the de-identification of QIs, which will significantly reduce the size of actual data for analysis purposes. The sensitive data identification research using QIs helps to minimize the information leakage of an individual entity.

The major objective of the research paper is to identify sensitive data and provide assurance of the security of the data. The research work is carried out based on medical healthcare data set. All the information provided by the patient is private but not always sensitive [16-18]. The authors have identified the sensitive data attributes given by the domain

expert in the cloud environment. The sensitive data of a patient cannot be disclosed as it hampers the privacy and security of the patient. The process designed by the authors is an automated process. The most exciting part of the research paper is that it considers the static and moving data request. The request may be generated from the same network as well as a global network. The designing of the Algo and the corresponding decision tables make the system architecture more robust and secure. The effectiveness of the entire architecture is validated in the medical healthcare system. The further enhancement of the research should consider the other domains like finance, the insurance sector, etc., to make the different systems integrated. The contribution of the paper is highlighted as follows.

- A novel Algo is proposed for the easy identification of sensitive data.
- Algo and corresponding decision tables are designed to make the system more robust.
- A graphical representation of the process flow is depicted as a part of the research paper.

The paper is organized as follow. The discussion on the related research is given in Section 2. In section 3, sensitive data identification methodologies are elaborated. The proposed model is also discussed in section 3. Security Section 4 highlights the security assurance of data. Section 5 highlights the result section, and the paper is concluded in Section 6.

2. Related Work

This section considers some of the valuable research carried out by scholars in the field of security in cloud computing. The sensitive data must not be revealed publicly. Multiple types of data are stored in a cloud-IoT-based computing model. Few data are sensitive, and few may not be sensitive [19-20]. If the sensitive data is not identified and protected by the cloud provider, then the privacy preservation of users will completely be destroyed. Transferring sensitive data between different internetworking systems also poses different challenges as data can be tampered with by illegitimate requests. Several researchers propose different schemes to tackle the major issues. Encryption of sensitive data may [21,22] be one of the solutions to protect the individual's data, but most cloud applications will not be able to decrypt the dataset easily. They provide their research direction with the help of encryption technology, but the researchers have not proposed any optimal encryption method to tackle the network congestion issue. Transferring the key from source to destination also consumes network bandwidth. The rapid growth of healthcare data in distributed computing environments finds a new challenge for privacy preservation techniques. For instance, cloud-based health applications grow exponentially. So new data are collected and added to the system. Identification of sensitive data and non-sensitive data over incremental data sets is a big issue. To cope with the problems Map-Reduce technique has been proposed by various researchers. It is a parallel data processing solution that segregates the sensitivity from the other non-sensitive data sets [23, 24]. They both suggest the parallel processing technique. Jain. P [23] suggests the Map-Reduce technique with the homogeneous data sets, whereas Meng. D [24] considers the entire research direction with the heterogeneous and generic data set only.

Some of the researchers suggest for K-anonymity model for the security assurance over sensitive data. The data is collected from different wearable devices, and they are easily identified. This could lead to a threat against the K-anonymity model. Sensitive data in transit may be more exposed as compared to data at rest. Data requires to be moved from one network to another in the transit phase. The intruder can easily hack the data and may alter the original data. Anonymization technique along with encryption methodology can be applied to overcome the problem. But the De-anonymization process at the destination end is time-consuming depending on the size of data [25, 26]. The researcher [25] considers the expected overhead of the system implementation to battle against the threats, whereas [26] does not provide any consideration regarding the same. For security assurance over sensitive data, some recommendations have been identified. Management of people and users' roles are to be strictly implemented in the cloud environment. When user applications move to the cloud, then the provider should assign their user identities to different access levels and this should reflect in their operational security policies. Different security Algo exist for the privacy and integrity of data. Sensitive data requires confidentiality and personal data needs privacy [27].

Zhao and Chen [28] have explained the security issues regarding the transferring of sensitive data to the centralized architecture. Users are not interested in moving their data to the cloud system. Different security assurance techniques are presented to enhance the security assurance technique on cloud system. N. Cao et al. [29, 30] have explained the different network attacks which disrupt the flow of data between internetworking systems. Chen and Zhao consider the data movement from both the local and global networks, but N.Cao considers all the attack scenarios based on intra network systems. The inter-networking attack scenario is not considered in their research paper. A. Gholmi et al. [31, 32] proposed the security framework for the cloud-based data transfer process. This platform supports to process of sensitive genomic data in the centralized system. The working model of the proposed framework is dependent on the cloud computing threat privacy model. To ensure security assurance two-factor authentication scheme and access control mechanism are also implemented.

3. Research Methodology

Identification of data attributes is very important with respect to the security assurance of data. A Standard de-identification process can prevent the breaching of sensitive data in distributive atmosphere. Leakage of information may destroy the reliability of the entire system [33, 34]. The quasi-identification techniques are utilized to make the main model of the research work. The quasi-Identifier technique can recognize attributes from the subset of the data set. Data collection and processing are done through distributed architecture. The automated security model enhances the integrity and reliability of the system architecture.

Fig.1. specifies that private data is segregated from public data. Thereafter, the private data is passed through the sensitive data identification process using a quasi-identifier. All data may not be sensitive. After the successful passing through the process, two sets of data are taken as output. One of them is sensitive data, and another one is non-sensitive data. Both the data can be archived for further processing.

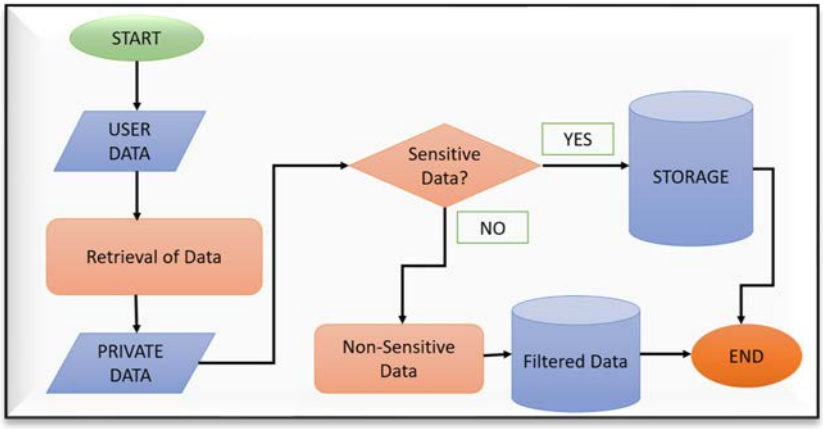

Fig.1. Process flow of sensitive data identification

3.1. Private Data Elimination from Personal Data

The primary objective is to highlight the conversion of personal data into private data. The research paper mainly considers medical data for the development of the experiment scenario. The patient's mobile number is his/her private data. Hence, private data cannot be disclosed to the external without the consent of a patient. Here, the storage location and type of data are different. The comments which don't contain names are stored in output set attribute T_s, and these are returned to the output data set.

Algo 1: Private Data Elimination from Personal Data

Inputs: Schema of 'M_M' sets of data

Outputs: Non-privacy data schema 'D_D'

```
1   Algo Estimating_Priv_Dat_From_Persnl_Data (Schema of 'M_M' sets of Data)
2   {
3   Foreach key L_i ϵ T_M where i=1…n
4   Input name of schema, foreign key, primary key and comments
5   Add it to T_D
6   Foreach attribute B_j ϵ T_D
7   if B_i! = O_k
8   Add it to U_s
9   Foreach attribute B_k ϵ U_s
10  if comments do not contain name
11  Add it to T_s
12  Return T_s
13  }
```

Fig.2. Private data elimination from personal data

Table 1. Representation of Notations and Symbols

Notation	Meaning
L_i	Key Attribute
T_M	Schema of M_M sets of data
T_D	non-privacy data Schema 'D'
B_j	Name of Attribute
O_k	Name of Primary key
U_s	Name of Temporary schema
T_s	Attributes that set the results

3.2. Identification of Sensitive Data from Private Data

The authors need to find out the attributes of sensitive data described the experts of the domain experts with the help of quasi-identifier. Domain experts provide attribute sensitivity scores. A domain expert is an expert person who has meaningful experience in the functional area [35, 36]. The domain experts and the corresponding expert systems for performing specific tasks are available, and these types of systems are appropriately defined as per the rules and regulations of the concerned department/section of the organization.

The sensitive attributes are defined as follows:

- Sensitive attribute: The sensitive attribute set, denoted by S_a, is the set of identifying attributes for the schema R (non-private data schema), i.e., $S_a \epsilon S_R$
- Rule: A rule condition R_x= R1⊞ R_2 R_2 is a condition with $R_1 \epsilon$ {schema name, attribute name, data type, unique key, primary key, foreign key, comments, and constraint.}. Now, this is consisting of disjunction and conjugations of rules using sensitivity score. σ[1,5], where σ permits the evaluation of the sensitiveness of an attribute that most satisfies the specified rule. The attribute sensitivity score is calculated based on the prescribed questionnaire format.

Algo 2: Sensitive data identity verification from private data

A process p(x), its initial pass p'(x) and primary approximation to the sensitive value ('ss') denoted by 'x'. A trivial threshold identifier is 'θ'and allowed passes maximum limit is 'N'. The schema of the sensitive data attribute is denoted by ss_m. Quasi-identifier technique is depicted by q. The threshold value indicates a checkpoint and stops the execution of the system passes at a predefined range. The calculation of threshold value is limited by the number of passes executed in the system architecture. The authors used a threshold value to compare the outcomes of *Pass 1* and *Pass 2* of the parallel and distributed processing. Further, schema attribute processing is dependent on the previously attained threshold value.

Inputs: Pass a private data set.

Outputs: Sensitive data schema generated.

```
1   Algo Generating_Sensitive _Schem _Fro _Privte_Dat (Private Data Set)
2   {
3   i=1
4   Foreach attribute Bi ϵTs
5   xAi=1 //The score of sensitivity
6   if p'(xi) is less than q(Ai)
7   { Move to step number 11; }
8   else
9   { Move to step number 16; }
10  xi+1= xi -p(xi)/p'(xi)
11  if |xi+1 -xi|<θ
12  Axi= Axi+1
13  i++;
14  if  (i ≤ N) { Move to step number 6; }
15  else { Move to step number 16; }
16  For each attribute Bi ϵTs
17  if xAi ≥ 5
18  Add Bi to ssm
19  }
```

Fig.3. Sensitive data identification from private data.

- In the proposed Algo - 2, the authors have scanned the resultant set of attributes received as outputs from Algo 1. In the starting phase, the sensitivity score of each attribute is set to 1. Here, $p'(x_i)$, which is the first pass of the given process $p(x)$, is checked with $q(A_i)$, and researchers have experienced two most probable outputs of the checking process. In a case if the obtained value is lower than A_i then calculation of $x_{i+1=}\ x_i - p(x_i)/p'(x_i)$ is done. After that, the difference of $x_{i+1} - x_i$ is calculated. If it is lesser than the small threshold value $\boldsymbol{\theta}$ then the obtained value of A_{xi} is incremented by 1. Further, the value of control variable 'i' should be increased until and unless it is lesser than the highest number of pass N., Further, if the $p'(x_i)$ is greater than the $q(A_i)$ then it is filtered out all the attributes from R_s whose sensitivity score is greater than 5 and store it in sensitive data schema set.
- The importance of Algo - 2 is to generate sensitive data set from the private data set. In this Algo, the identification of sensitive data is completely based on the sensitivity scores provided by the expert system of the domain. The described A*lgo-2* will consider only those attributes as sensitive whose sensitivity is score is greater than or equal to 5. The Algo -2 effectively finds out the sensitive attribute based on values provided by the domain experts/expert systems.

3.3. Mathematical Modelling of Proposed System

The authors have designed SISA called as sensitive-data identification and security assurance technique using mathematics based modelling. This mathematics based model is an intangible representation of SISA and it is described in below subsections:

A. Structure of SISA components

Composition of different components of SISA model can be represented by Eq. (1).

$$SISA = \{W, \sum\ ,\ V, Q_{f,} R_{f\ ,}\} \tag{1}$$

The symbolic representation of Eq. (1) is described below:

W→ finite set of states which are non-empty.
Σ → non-empty finite set of known input matrices to a part of SISA
V →non-empty finite set of received outputs of SISA
Q_f → SISA's state transformation function representing different states.
R_f →The received output function.
P_s → The preparatory state. It is the starting phase of SISA, and this can be treated as a subpart of 'W'. i.e., $V_s \boldsymbol{\epsilon}$ W.

Each state can primarily be recognized as an individual processing unit whose authenticity is solely dependent on the input data obtained from different IoT related devices.

B. Interrelation of SISA parts

The SISA components are described below:

- States set ('W'): The inputs given and the outputs obtained from the individual states of SISA are identified.
- Input sets / output sets ('Σ'/ 'V'): Alphabet set 'Σ' is the combination of all input to the proposed model at subsequent states.

  Let P $\boldsymbol{\epsilon} \sum$,
  $\Longrightarrow$ $P_1 \times_1 \times_1$is the easiest element$\sum$, and it is a real number.
  if
  P $\boldsymbol{\epsilon} \sum$, then [P]=$p_{l,m,n}$ =p(l, m, n) $\forall\ l, m, n \boldsymbol{\epsilon}$ $\{Z^+ - 0\}$.
  V is the output alphabet set of SISA. The below equation is true for 'V'.
  Σ=$\{y | y_{i \times j \times k}$ is a three-dimensional matrix$\}$

C. The Representation of 'Q_f'

The 'Q_f'can be represented by Eq. (2) given below:

$$Q_f : \overset{*}{\sum} \times W \rightarrow W \tag{2}$$

In Eq. (2), $\sum^*$ is the set of the basics of Σ.

D. The output function (R_f)

The 'R_f' is represented by Eq.(3).

$$R_f: \overset{*}{\sum} \times W \rightarrow V^* \tag{3}$$

Here, V^* is the representation of all processes obtained from the essentials of 'V'
The Eq. (4) represents the true value of each tuple.

$$W= \{w_0, w_1, w_2, w_3\}$$

$$\Sigma= \{PE_D, PR_D, SS_D, DS_D\}$$

$$V= \{PE_D, PR_D, SS_D, DS_D\}$$

$$Q_f: W \times^* \rightarrow W \text{ and } R_f: W \times^* \rightarrow V^* \tag{4}$$

In Eq. (4) PE_D is the patient's private information/data of a hospital's patient, and PR_D is the private data received through data extraction. Now, if 'p' $\epsilon \sum^2$ and 'w' ϵ W, then Q_f(w, p) validates the processing completed on 'p' $\epsilon \sum^2$, at state wϵ W. Here, 'w' represents an Algo.

4. Security Assurance over Sensitive Data

The research paper focuses on the designing of security assurance techniques in a novel way. The security system is a completely intelligent-based system and works optimally without human intervention. The identification of different threats is automated, and subsequent information is stored in the suspicious database system. Therefore, the system administrator can easily trace unauthorized data requests. As soon as the request is forwarded from a different system, it will be validated internally. The positive response will allow the request to access data at a certain point of time. A negative response will discard the request information, and an entry will be saved to the suspicious database [37-39].

Algo 3: Algo for getting static sensitive data from local net

Input: Passed Parameters; ***Output:*** Decision of static accessing the sensitive data

```
1   Algo Acces_Sati_Sensi_Dat_Loc_Network (Q1, Q2, Q3, Q4, Q5)
2   {// True is represented by '1' and False is represented by '0'
3   if ((Q1== 1) && (Q2== 1))

4   Print "DoS to the request";
5   else if
6   ((Q2== 1) && (Q5== 1))
7   Print "IP is blocked to send request";
8   else if
9   ((Q2== 1) && (Q3== 1))
10  Print "Reported the net administrator";
11  else if
12  ((Q3== 1) && (Q4== 1))
13  Print "Approval of access is awaited";
14  else
15  Print "Access permission given";
16  }
```

Fig.4. Access of static sensitive data obtained from the local network

4.1. Request for Using Static Sensitive Data from a Local Network (1st Case)

The requirement is generated from a network where sensitive data is located. The system administrator will manage the suspicious IP addresses to check the authenticity of the request. The successfully validated request can get entry to the network system. Security assurance technique for case 1 is as follows.

Q_1: Demand is arriving from the suspicious IP address
Q_2: Demand harms the current network.
Q_3: Demandaccesses any sensitive data without proper approval
Q_4: Demand tries to access data after time elapses
Q_5: Demand is made for the degradation of the network

4.2. Request for Using Static Sensitive Data through the Global Net (2nd Case)

The origination of the request is completely different from the location of the sensitive data. The system administrator does not have sufficient knowledge about the authenticity of the request. The information will be collected from the adjacent networks. The positive response will be considered for the entry of the request. Priority of the request is also a major consideration part of this technique. Local requests will be more privileged than global requests. Case 2 parameters are explained as follows.

R_1: Demand has come from a suspected IP.
R_2: Demand has no positive impact on nearby networks
R_3: Demand has history of using sensitive data of current network.
R_4: Demand is trying to access data beyond specified time.
R_5: Demand is not authorized.

Algo 4: Algo for getting static sensitive data from global net

Inputs: Passed parameters; ***Outputs:*** Decision to access the sensitive data

```
1   Algo Acces_Satic_Sensitive_Dat_Glob_Network (R1, R2, R3, R4,R5)
2   { // True is represented by '1' and False is represented by '0'
3   if ((R1== 1) && (R5== 1))
4   Print "Denial of service";
5   else if
6   ((R2== 1) && (R5== 1))
7   Print "The IP is blocked";
8   else if
9   ((R2== 1) && (R3== 1))
10  Print "Net admin should be reported";
11  else if
12  ((R3== 1) && (R4== 1)
13  Print "Wait";
14  else
15  Print "Permission given to access the data";
16  }
```

Fig.5. Access to static sensitive data from the global network

4.3. Request for Using Locally Moving Sensitive Data through a Local Net (3rd Case)

The demand is originated from the same location of sensitive data, but data is moving in this regard. The position of sensitive data is always changing. The access right of the specific node is also an important factor before giving the approval to access the data. The access right is time bound and also validated through the suspicious database. The considered parameters are explained below.

S_1: Demand has come from suspected IP to access moving sensitive data
S_2: Demand data is busy.
S_3: Sensitive data access right given w.r.t. a node.
S_4: Demand is trying to access data after its allotted time.
S_5: Demand linked with damaging entities.

4.4. Request for Getting Globally Moving Sensitive Data Using Global Net (Case 4)

The origin of the request and sensitive data are at different locations as well as sensitive data is not static. The position of the sensitive data and its corresponding node access rights are checked by the system administrator. The successful validation of the request will get a specific time slot for accessing sensitive data. The request permission will be deactivated after the time elapses.

T_1: Demand has come from suspected IP to access moving sensitive data.

T_2: Demand data busy.
T_3: Demand for a specific node
T_4: Demand is trying to beyond its given time without any prior information.
T_5: Demand is linked with unauthentic elements.

Algo 5: Algo for getting moving sensitive data from local net

Inputs: Passed Parameters; ***Output:*** Final conclusion to access the sensitive data

```
1   Algo Acces_Mov_Sensit_Dat_Loc_Network (S1, S2, S3, S4,S5)
2   { // True is represented by '1' and False is represented by '0'
3   if ((S1== 1) && (S5== 1))
4   Print "Denial of service";
5   else if
6   ((S2== 1) && (S5== 1))
7   Print "IP Address blocked";
8   else if
9   ((S2== 1) && (S3== 1))
10  Printf  "net administrator reported";
11  else if
12  ((S3== 1) && (S4== 1)
13  Print "Wait";
14  else
15  Print "Access Permission given";
16  }
```

Fig.6. Access to moving sensitive data from the local network

Algo 6: Algo for getting moving sensitive data from global net

Inputs: Passed Parameters; ***Output:*** Decision to access the sensitive data

```
1   Algo Acces_Mov_Sensit_Dat_Glob_Net (T1, T2, T3, T4, T5)

2   { // True is represented by '1' and False is represented by '0'
3   if ((T1==1) && (T5== 1))
4   Print "Denial of service";
5   else if
6   ((T2== 1) && (T5== 1))
7   Print "IP address is blocked";
8   else if
9   ((T2== 1) && (T3== 1))
10  Print "net administrator reported";
11  else if
12  ((T3== 1) && (T4== 1)
13  print "Wait";
14  else
15  Print "Access permission given";
16  }
```

Fig.7. Access to moving sensitive data from global net

5. Analysis of Results and Discussions

The basic part of this paper mainly considers the data identification process and provides a complete comparative analysis of the existing processes. The next part clearly highlights the sensitive data security assurance methodologies for the cloud-based system. The decision tables are highlighted for easy identification of the threats.

5.1. Analysis of the Sensitive Data Identification Algo

From the above attributes, patient ID, Patient Name, Patient Last Name, and Patient date of birth are personal data attributes. Patient phone numbers can be treated as a private attribute. Domain expert provides the sensitive data

attributes from a given set of data, and the Algo extracts it from heterogeneous sources of data. In India, generally, four types of treatments are available, i.e., Allopathic medicine, Ayurvedic medicine, Homeopathic medicine, and Unani medicine. The patient treatment information must be limited to these four sectors. The specified attributes are stored with various names and their databases are completely different. The data source may be a spreadsheet, word document, open-source databases like MySQL, PostgreSQL, proprietary databases like MS Access, MS SQL Server, oracle, or scam document like pdf. Finding a specific attribute among different heterogeneous databases is a big issue [40]. The proposed Algo 2 of Fig**.** 3 is used to find out the similar attributes/comment from the huge dataset. As an example, the identification of sex information in the medical domain is not very easy. In different databases, this information is saved in gender, sex, Lingo, etc. To get the information from different databases domain expert will specify some attributes/comments. The proposed Algo will search the similar matching from the databases. It is observed that doctors specify different codes in their prescriptions [41].

5.2. Comparison of the Proposed Sensitivity Algo with the Existing Algo

Presently sensitive data identification is a booming research topic. Several researchers are carried out their research in this field. They have identified different parameters to explain and justify their Algo. In this paper, we have identified a detection technique, De-identification technique, basic methodology, domain, and size of data set.

Table 2. Comparative study of the existing Algo with the proposed sensitivity Algo

	Authors Detail	Approach Suggested	De-Identification	Basic Methodology	Restrictions	Domain	Large Dataset	Sensitivity Score
1	C.Du Mouza [42]	Sensitive Information	None	Semantic rules with NLP, semantic Modelling	Applicable for single attribute	Generic	Not Established	None
2	S. Lodha [43,44]	Quasi Identifiers	Probabilistic generalization and suppression	Probabilistic	It May is not accurate for unclean data	Generic	Not Established	None
3	D. Motwani [45]	Quasi Identifiers	Masking using a random sampling technique	Random sampling	Sacrifices accuracy by allowing approximate answers	Generic	Established	None
4	A. Omer[46]	Minimal set of attributes that identify maximum records	Substitution and generalization	Generalization hierarchies and decomposition	Considers only the minimal set of attributes and not all possibilities	Generic	Not Established	None
5	N. Mohammed [47]	Not suggested	Distributed anonymization	Probabilistic, Distributed Anonymization	No detection methodology suggested	Healthcare	Established	None
6	F. Liu [48,49]	Quasi identifiers technique	LKC privacy model	The greedy Algo requires multiple scanning of the table	Not accurate for unclean data	Generic	Not established	None
7	Z. Brakerski [50,51]	Not suggested	Distributed anonymization technique	Random sampling	No detection methodology suggested	Healthcare	Established	None
8	Proposed Algo	Sensitive data identification	Distributed anonymization technique	Probabilistic, random sampling, and multiple scanning of the table	It works in unclean data also	Healthcare	Established in a large dataset	Yes

The identification process retrieves types of data. Data is not always sensitive in nature. So, this technique indicates the identification of sensitive or non-sensitive data. De-identification of data is also very important. Otherwise, the privacy of the data may be destroyed. Different data de-identification techniques are available. Substitution and generalization techniques are also useful when data segregation is difficult. The identification Algo is completely dependent on the different rules. Rules can be established by the NLP technique or distributed probabilistic anonymization. Framing of rules is possible using the greedy Algo based technique, which requires multiple scans of the table. In the distributed system, all the datasets may not be in the same place. So multiple scanning of tables is obligatory. The existing Algo mainly work in the nonspecific domain. The approaches of the active processes based on the changes in the domain are not established. The size of the dataset is a very important parameter in this field. A few Algo work in a large dataset, but many of them fail to provide a complete vision in a large dataset.

In this paper, the authors have identified a novel parameter named sensitivity score, and a range of sensitivity scores is given by the experts of the domain.

5.3. Security Assurance Analysis of Sensitive Data

Section 4 completely presents the different cases for security assurance over sensitive data. This section provides the decision table based on each case and through which net administrators can easily validate the authenticity of the request. The following part provides the decision table for each of the cases [52-54].

A. Decision table for accessing static sensitive data from the local net

In this case, the request is coming from the same net where the sensitive data is stored. The net administrator does not directly allow the request to access the data until and unless the request is found to be legitimate. To check the status of the request net administrator will set a few parameters, which are placed in the decision table. The complete decision table is given in Table3

Table 3. The request is getting static sensitive data from the local network

	Name of the Parameter	**Status of the Action ('✓'/'X')** **'✓': allow 'X': Deny**			
1.	The request tries to access any sensitive data without authorization.	X	X	✓	✓
2.	Demand is trying to access data after allocated time.	X	X	X	✓
3.	Demand is arriving from the unauthenticated IP address	✓	X	✓	X
4.	Demand harms the current network.	✓	✓	✓	X
5.	Demand is not related to authorized entity.	X	✓	X	X
Action Statements					
1.	DoS	✓	X	X	X
2.	Inform the Net admin	X	X	✓	X
3.	Wait for approval of access	X	X	X	✓
4.	Obstruct the IP for sending the request	X	✓	X	X

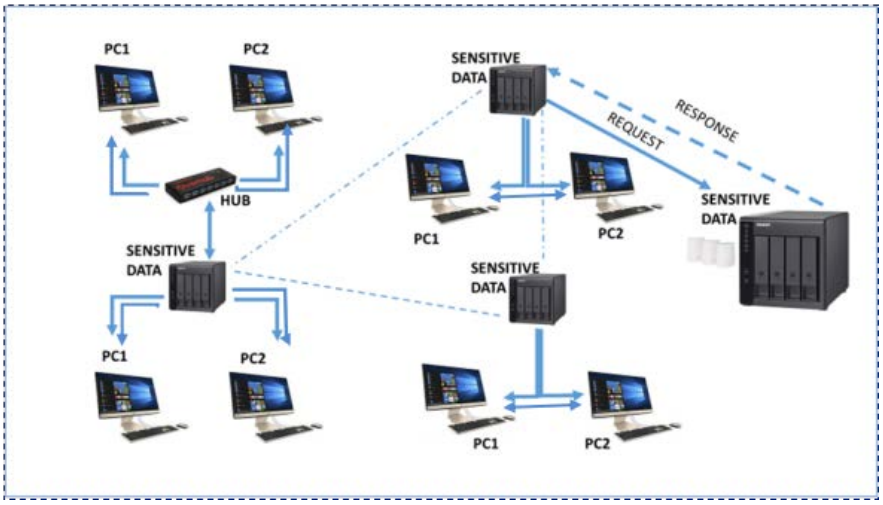

Fig.8. Request to get static sensitive data from local net

Fig.8 represents a complete net where sensitive data is stored in the server. Sensitive data is static. The system connected to LAN-2 is sending a request to access te data. The request cannot be granted instantly. It will be validated based on Algo- 3 of Fig.4 and Table 3. If the validation is successful, then a request will be granted to access the data. Otherwise, it will be rejected, and the requested IP will be saved into the suspicious IP database.

B. Decision table for accessing static sensitive data from the global net

The request and sensitive data are located in different networks. The identity of the request is completely unknown to the net administrator. So, the administrator will not directly allow the request to enter the network. The administrator will check the status of the request from the decision table. The complete decision table, along with the parameters, is explained in table 4.

Table 4. The request is accessing static sensitive data through a global network

	Name of the Parameter	Status of the Action (‘✓’/ ‘X’) ‘✓: Allow ‘X’: Deny			
1.	Demand has bad history.	X	X	✓	✓
2.	The request is trying to access data after its fixed time without a pre-request.	X	X	X	✓
3.	Demand has no positive impact on its adjacent networks.	X	✓	✓	X
4.	Demand is arriving from the suspicious IP address	✓	X	X	X
5.	Demand is not related to authorized entity.	✓	✓	X	X
Action Statements					
1.	DoS	✓	X	X	X
2.	Inform the Net admin	X	X	✓	X
3.	Wait for approval of access	X	X	X	✓
4.	Obstruct the IP for sending the request	X	✓	X	X

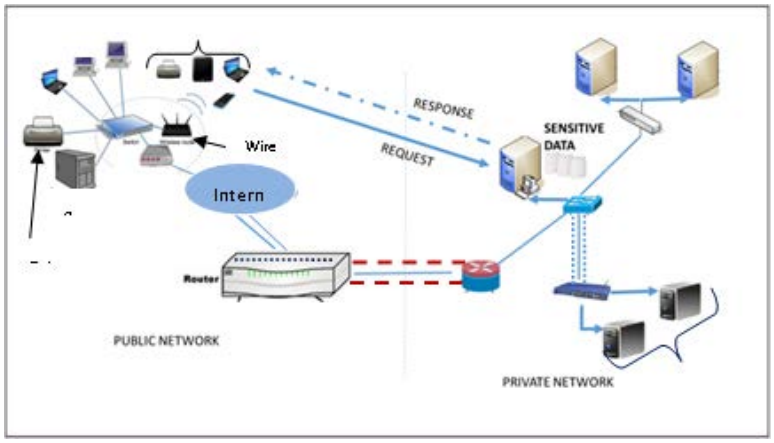

Fig.9. Request to access static / moving sensitive data from outside of the network

Table 5. The request is accessing locally moving sensitive data through a local network

	Name of the Parameter	Status of the Action (‘✓’/ ‘X’) ‘✓’:Allow ‘X’: Deny			
1.	Access right of the sensitive data for a particular node	X	X	✓	✓
2.	The request is trying to access data after its fixed time without a pre-request.	X	X	X	✓
3.	Demand is from suspected IP.	✓	X	X	X
4.	Sensitive data is busy.	X	✓	✓	X
5.	Demand is not related to any authorized entity	✓	✓	X	X
Action Statements					
1.	DoS	✓	X	X	X
2.	Inform the Net admin	X	X	✓	X
3.	Wait for approval of access	X	X	X	✓
4.	Obstruct the IP for sending the request	X	✓	X	X

Fig. 9 clearly shows a complete net where sensitive data is stored in the server. Sensitive data is static. The server is connected to LAN-1 under network 1. The request is coming from net 2 to access the sensitive data. The request cannot be granted instantly. It will be validated based on the Algo -4 of Fig.5. and Table 4. If the validation is successful, then a request will be granted to access the data. Otherwise, it will be rejected, and the requested IP will be saved into the suspicious IP database.

C. Decision table for accessing moving sensitive data from the local net

The sensitive data is not always static. It can be moving from one node to another inside the network. The net

administrator will identify the parameters to create the decision table. If the request is found to be legitimate, then it will be able to access sensitive data. The complete decision table, along with the parameters, is explained in Table 5.

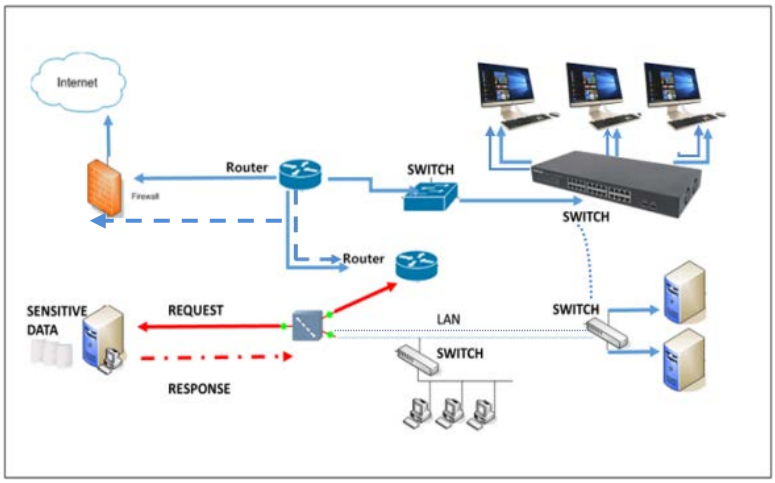

Fig.10. Request to access moving sensitive data from the local network

Fig.10 clearly shows the complete diagram where sensitive data is not static. It is moving from one node to another. The request is trying to access the data from the same network. The request cannot be granted instantly. It will be checked based on the rules framed in Algo 5 of Fig.6 and the corresponding Table5. If it is validated successfully, then data can be accessed. Otherwise, it is rejected [55, 56].

D. Decision table for globally moving sensitive data from the global net

The movement of sensitive data cannot be restricted to a single net only. It can move from one net to another. Request can also be forwarded from the global network. In this case, the net administrator will not grant the request to access sensitive data instantly. The request will be validated based on Algo - 6 of Fig. 7 and Table 6. If the validation is successful, then a request will be permitted to access the data**.** The complete decision table and other parameters are explained in Table 6.

Table 6. The request is accessing globally moving sensitive data through a global network

	Name of the Parameter	**Status of the Action ('✓'/ 'X') '✓': allow 'X': Deny**			
1.	Sensitive data access right requested for a node.	X	X	✓	✓
2.	The request is trying to access data after its fixed time without a pre-request.	X	X	X	✓
3.	Demand is from suspicious IP.	✓	X	X	X
4.	Sensitive data is busy.	X	✓	✓	X
5.	Demand is not related to an authorized entity.	✓	✓	X	X
Action Statements					
1.	DoS	✓	X	X	X
2.	Inform the Net admin	X	X	✓	X
3.	Approval awaited	X	X	X	✓
4.	Obstruct the IP for sending the request	X	✓	X	X

Fig.11 clearly shows the complete diagram where sensitive data is not static. It is moving from one net to another network. The request is trying to access data from the outside of the network. The request may not be granted immediately and it will be verified based on algo-6 of Fig.7 and Table6. If it is validated successfully, then data can be accessed. Otherwise, it is rejected.

5.4. Implementation Atmosphere

The proposed algo of security assurance over sensitive data have been implemented in the Cloud Simulation technique. The experiment considers six cloudlets, virtual machines of 10^6, and memory is 1024KB in dual-core configuration. The experiment is considered for changing values of passes from 100 to 300.

5.5. Implementation Level Analysis for Static Sensitive Data

In the previous sections, Table3and Table 4 indicates the rule to access static sensitive data based on the local request and global request. After combining both the tables and excluding similar parameters we used to receive the

following parameters:

P_1: Demand is not from an authorized IP.
P_2: Demand is harming current network.
P_3: Demand harms the nearby net.
P_4: Demand tries to access sensitive data without authorization
P_5: Request is trying to access sensitive data after its allocated time without any prior information.
P_6: Demand is not related to authorized entities.

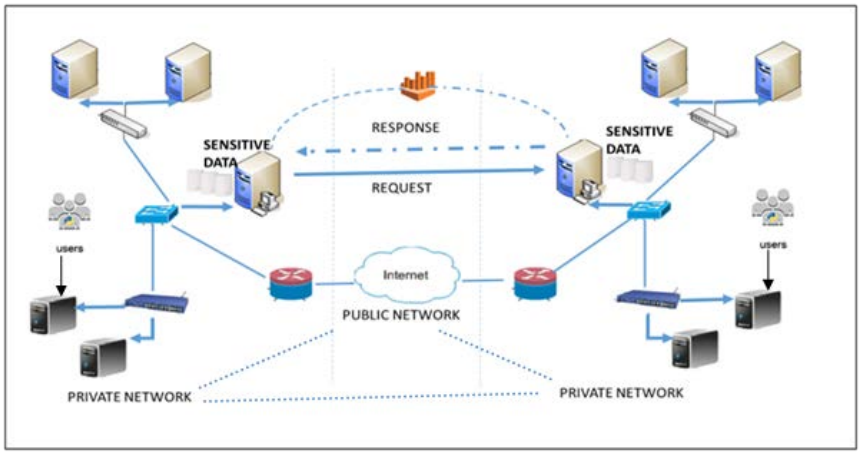

Fig.11. Requesting to use moving data from the local / global network

When a request comes to access static sensitive data, we have analyzed the impact of the above-mentioned parameters. We have executed the simulators and received the changes to the scenario during each pass. The entire scenario is depicted ted using 2-D and 3-D graphs. In the graph, the 'x-axis denotes the request parameters, and the 'y' axis specifies the normalized throughput. The number of requests handled by each of the cloudlets per unit of time is the actual throughput [57].

It was observed by researchers that P2 has the lowest number of hits. This changing scenario of the parameters indicates that based on the requests, the described system provides a restriction on accessing illegitimate data. Otherwise, a change in parameters would not be required. The different scenarios of change of parameters for *1st Pass*, *2nd Pass*, and 3rd *Pass* are depicted in Fig. 12 using a 2-D graphical representation.

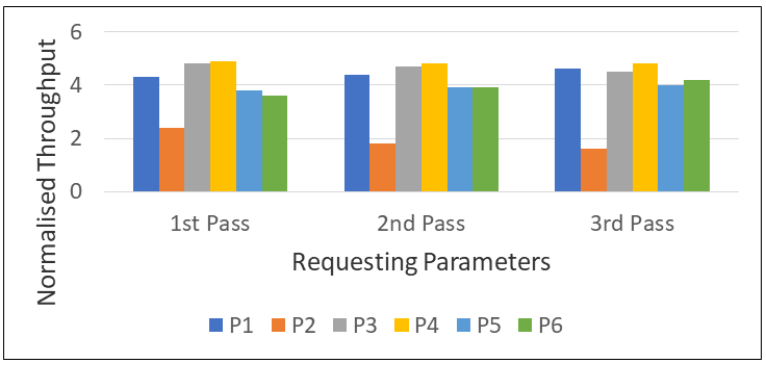

Fig.12. 2-D representation of static data request

Fig.13 represents the 3-D graph of static data request, which shows that the parameters are getting changed based on changes in passes. This request can't be granted instantly.

5.6. Implementation Level Analysis of Moving Sensitive Data

After combining both the tables and excluding similar parameters the authors received values related to the following parameters:

P'_1: Demand is coming from a suspicious IP.
P'_2: Sensitive data is busy.
P'_3: Demand has no authorization to access sensitive data
P'_4: Access right not given to any node for sensitive data.
P'_5: Demand is trying to access data after its allocated time without any prior information.

P'$_6$: Demand is not related to any authorized entity.

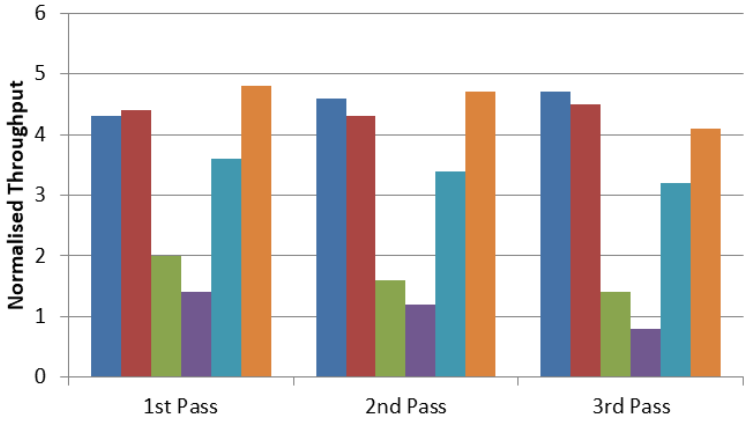

Fig.13. 3-D representation of requests of static data.

When a request comes to access moving sensitive data from any site, the researchers described the impact of the parameters from *P'$_1$* to *P'$_6$*. The researchers executed the simulator, and with each pass, they received the changes in scenarios. The entire scenario is depicted in Fig. 14 and fig. 15 using 2-dimensional and 3-dimensional graphs. Further, the throughputs were considered to be between six cloudlets, so the parameters were also changing based on passes from pass 1 to pass 3. The 2-dimensional graphical representation of the request to access the moving sensitive data is presented in Fig.14.

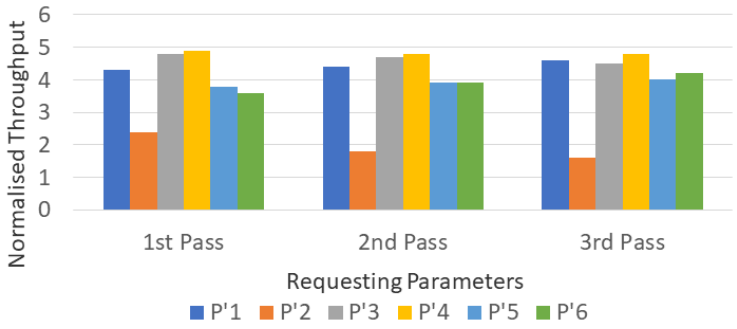

Fig.14. 2-D depiction of moving sensitive data requests.

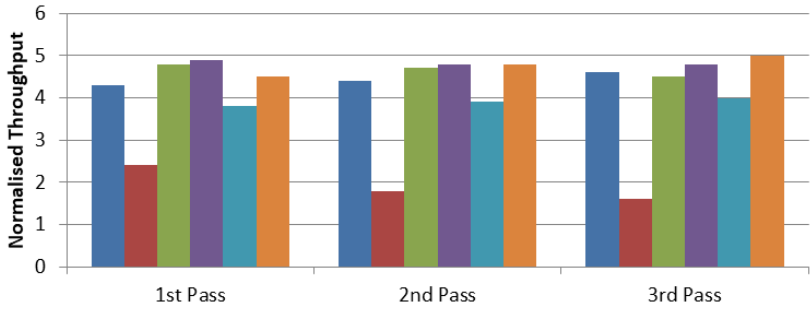

Fig.15. 3-D depiction of moving data requests.

Fig. 15 is a diagrammatical representation of moving sensitive data in the form of a 3-dimensional graph. It was observed by the authors that the parameters were changing for normalized throughput after the happening of pass change from 100 to 300. The requests were observed first, and then permission was given by the proposed system if the request was found to be legitimate. The complete information about the unauthenticated requests was saved to the internal database of the automatic net system.

6. Conclusion

In the current era, sensitive data integrated distributed computing environment are getting a lot of attraction. Every day, millions of dollars are spent by companies on the identification and security of sensitive data, and still, the anxiety continues. The challenges of secure data identification and their security assurance-related issues are knocking rapidly because these data are being shared and used frequently in a cloud computing environment to execute specific tasks.

The authors believe that if sensitive data are not identified and protected from intruders, then the privacy of the internetworked users will be destroyed. The authors have presented novel techniques of sensitive data identification, analysis, and security assurance mechanism in this paper. To assure the security of sensitive data, the authors have used various clouds and IoT-compatible environments. The authors have presented a mathematical model to describe the interconnections between the components of the proposed model, where each component takes some inputs, performs the processing tasks, and produces the outputs. The decision tables of security assurance cases of this research work help make decisions while giving permission to access the sensitive data. The outputs of one component of the proposed method may be given to another component as the inputs in the cloud-IoT integrated computing environment. The proposed sensitive data identification Algo takes care of finding the matches of attributes that are specified by the experts of the domain, and these attributes are derived from different data sources of clouds. The security assurance Algo, and their corresponding decision tables will help the internetworked administrators to find out the legitimate request for permitting to access sensitive data.

References

[1] Rajendra P., Harsha D., Chirag M., "Designing an efficient security framework for detecting intrusions in a virtual net of cloud computing", Computers & Security, Volume 85, pp. 402-422, August 2019

[2] Ado Adamou Abba , Olgaengni N., Chafiq T., Ousmane T., Alidou M., Abdelhak M.G., "Enabl priva and secu in Cloud of Things: Archit, apps, security & priv challeng", Appl Comp & Info, pp. 1-23, 2019.

[3] Ahmed Kayed, Suha Omar," Periodical Key change for cloud mutable security protocol," Microprocessor and Microsystems, Vol.69, pp. 152-158, 2019

[4] Reyhaneh Rabaninejadet al.," Comments on a lightweight cloud auditing scheme: Security analysis and improvement," Journal of Net and Comp Apps, Vol.139, pp.49-56, (2019)

[5] Poornima M. C., Mahabaleshwar S. K., "Security and Privacy in IoT: A Survey," Wireless Personal Communications, Vol. 115, pp. 1668-1693, 2020.

[6] Jin L. et al., "Security and privacy in IoT communication," Annals of Telecommunications, Vol.74, pp. 373-374 (Editorial),

[7] Yuqing M., "A Data Security Storage Method for IoT Under Hadoop Cloud Computing Platform," Int. J. of Wireless Info Nets, Vol.26, pp.152-157, 2019.

[8] Pankaj K., Lokesh C., "A sec authentic scheme for IoT app in the smart home", Peer-to-peer Net & Apps., Volume 14, pp. 420-438, 2021.

[9] Sutrala, A.K et al., "Secure anonymity-preserving password-based user authentication and session key agreement scheme for telecare medicine information systems," Comput. Methods Programs Biomed. 135, 167–185, 2016.

[10] A. Dorri, S. Kanhere, R. Jurdak, "Blockchain for IoT security and privacy: The case study of a smart home," IEEE Int. Conf on Perv Comp and Comm Wkshop, pp. 1-6, 2017.

[11] Casas-R. Et al., "A summary of k-degree anonymous methods for privacy-preserving on networks," Adv Res in Data Priv, 567, pp231-250, 2015.

[12] Amiri F. Et al., "Hierarchical anonymization Algo against background knowledge attack in data releasing," Knowl-bsd Sys, Volume.101, pp. 71-89, 2016.

[13] Nayahi J.J.V., Kavitha V. "Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop", Fut Gen Comp Sys, Int J. of Sc., Volume.74, pp393-408, 2017.

[14] Sarkar, M. et al., "Configuring a trusted cloud service model for smart city exploration using hybrid intelligence," Int. J. of Amb Comp and Intell., Vol. 8, No. 3, 1-21, 2019.

[15] Palanisamy B., Liu L., "Privacy-preserving Data Publishing in the Cloud: A Multi-level Utility Controlled Approach," Proc. of 8th Int Conf on Cloud Comp., pp. 130-137, 2015.

[16] Wang J. et al., "RPRep: A Robust andPrivacy-Preserving Reputation Management Scheme for pseudonym-Enabled VANETs," Int. J. of Distr. Sens Net., Volume12, No. 3, pp. 1-15, 2016.

[17] Terrovitis M. et al., "Local Suppression and Splitting Techniques for Privacy-Preserving Publication of Trajectories," IEEE Transactions on Knowledge and data Engineering, Vol. 29, No. 7, pp. 1466-1479, 2017.

[18] Hossain M.M. et al., "Towards an analysis of security issues, challenges, and open problems in the internet of things", IEEE World Cong., pp. 21–28, 2015.

[19] Dey, N. et al., (Eds.) "Internet of things and big data analytics toward next-generation intelligence," pp. 3-549, 2018.

[20] Mohan A., "Cyber security for personal medical devices internet of things,", Distri. Comp. in Sens Sys., Int Conf. of IEEE, pp. 372–374, 2014.

[21] Yoon S. Et al., "Security issues on smart-home in IoT environment," in Computer Science and its Applications. Springer, pp. 691–696, 2015.

[22] Abdur M.R. et al., "Digital image security: Fusion of encryption, steganography and watermarking", Int. J. of Adv. Comp. Sc. & Apps, Volume 8, No. 5, 2017.

[23] Jain, P.et al., "Big data privacy: a technological perspective and review", J. of Big Data, Volume 3, (2016).

[24] Meng, D. "Data security in cloud computing", Int Conf on Comp Sc & Educat, pp. 810–813, 2013.

[25] Shaikh, R., and Sasikumar, M. "Data classification for achieving security in cloud computing," Proc. Comp. Scien., Volume. 45, 493–498, 2015.

[26] Shuijing, H. "Data security: the challenges of cloud computing,", Int Conf on Measur Tech and Mech. Automat, pp. 203–206. 2014.

[27] Singh, V.K. and Singh, T. "Present data security issues and their resolving technique in cloud computing," International journal of Science & Technology., Volume. 1, pp. 1–6 (2016)

[28] Chen, D and Zhao, H. "Data security and privacy protection issues in cloud computing," In Int. Conf. on Computer Science and Electronics Engineering, pp. 647–651, 2012.

[29] Cao N et al., “Privacy-preserving multi-keyword ranked search over encrypted cloud data," IEEE Transactions on Parallel and Distributed Systems, Volume 25, No. 1, pp. 222–233, 2014.
[30] Lauter K et al., "Private computation on encrypted genomic data," Technical Reports. MSR-TR-2014-93, June 2014.
[31] Gholami A. and Laure E., “Advanced cloud privacy threat Modelling," International Conference on Software Engineering and Applications, pp. 1-11, 2015.
[32] A. Gholami et al., "Privacy-preservation for publishing sample availability data with personal identifiers," Journal of Medical and Bioengineering, Volume 4, pp. 117–125, April 2014.
[33] Tari Z, et al., “Secur and priv in cloud comp: Vision, trends, and challngs”, Cloud Comp., 2(2), pp. 30-38, 2015.
[34] Singh A., Chatterjee K., “Cloud security issues and challenges: A survey”, J of Net and Comp Apps., Volume. 79, pp. 88-115, 2017.
[35] Leng C, Yu H, Wang J, Huang J., “Securing personal health records in the cloud by enforcing sticky policies," Indonesian J. of Elect Engg and Comp Sc., Volume 11, No. 4, pp. 2200-2208, 2013.
[36] Zhou L. et al., "Integrat trust with cryptograph role-based acces contrl for sec cloud data storge," in Trust, Sec. and Priv in Comp and Comm., 12th Int Conf on, pp. 560–569, 2013.
[37] Sendor J. et al., "Platform level support for authorization in cloud services with oauth 2," Proc of Int Conf on Cloud Engg, pp. 458–465, 2014.
[38] Leandro M. A. et al., “Multitenancy authorization system with federated identity for cloud-based environments using shibboleth,”, Proc. of Int Conf on Netwks, pp. 88–93, 2012.
[39] Coppolino L. et al., “Cloud security: Emerging threats and current solutions," Comp. & Elect. Engg., Volume 59, pp. 126-140, 2017.
[40] Chen W. et al., “Outsourced privacy-preserving decision tree classification service over encrypted data," J. of Info Sec & Apps, Volume 53, August 2020.
[41] Hu Xiong, Yan Wu, ChunhuaSu, Kuo-hui Yeh, "A Secure and efficient certificates batch verification scheme with invalid signature identification for the internet of things," J. of Info Secur and Apps, Vol. 53, No. 8, pp. 1-12, August 2020.
[42] Du M. et al., “Towards an automatic detection of sensitive information in a database,” 2nd Int. Conf. on Adv in DBs, Knowldg, and Data Apps, pp. 247–252, 2010.
[43] Lodha S. and Thomas D., “Probabilistic anonymity,” in Priv, Secy, and Trust in KDD, Intl Workshop, pp. 56–79, 2007.
[44] Wang F., Han S. "Probabilistic Graphical Models: theory and technology," Tsinghua Univ. Press, pp. 1-1270, 2015.
[45] Shirudkar K., & Motwani D., “Big-Data Security,” Intl J of Advncd res in Comp Sc and S/w Engg Res., Volume 5, No. 3, pp. 1100–1109, 2015.
[46] Omer A. M. et al., "Simple and effective method for selecting quasi-identifier," J. of Theor and Appld Info Tech., Volume 89, No. 2, pp. 512–517, 2016.
[47] Mohammed N. et al., “Centralized and Distributed Anonymization for High-Dimensional Healthcare Data,” ACM Trans on Knowldg Discov from Data, Volume. 4, No. 4, pp. 1–33, 2010.
[48] Liu F. et al., “Privacy-Preserving Scanning of Big Content for Sensitive Data Exposure with MapReduce,”, Proc. of 5th ACM Conf on Data and Appl Sec and Priv., pp. 195–206, 2015.
[49] Baek J. et al., "A secure cloud computing-based framework for big data information management of the smart grid," Cloud Comp., IEEE Trans. on, Volume. 3, Issue. 2, pp. 233-244, 2014.
[50] Brakerski Z., Gentry C., and Vaikuntanathan V., “Fully Homomorphic Encryption without Bootstrapping,” In Proc. Innov. in Theoretical CS, pp. 309–325, 2012.
[51] Brakerski Z., “Fully homomorphic encryption without modulus switching from classical GapSVP,” Lect. Notes in Computer Science Volume 7417, pp. 868–886, 2012.
[52] Kamta N. M., "A Proficient Mechanism for Cloud Security Supervision in Distributed Environment," International Journal of Computer Network and Information Security, Volume 12, No. 6, pp. 57-77, 2020.
[53] Kamta N. M. et al., "Veins Based Personal Identification Systems: A Review," International Journal of Intelligent Systems and Applications, Volume 8, No. 10, pp. 68-85, 2016.
[54] Kamta N.M., and Anupam A., "A Soft Computing Technique for Improving the Fidelity of Thumbprint based Identification Systems,", International Journal of Intelligent Systems and Applications, Volume 8, Number 7, pp. 14-27, 2016.
[55] Kamta N.M. et al., "Palatal Patterns Based RGB Technique for Personal Identification,", International Journal of Image, Graphics and Signal Processing, Volume 7, Issue 10, pp. 60-77, 2015.
[56] Kamta N.M., and Kanderp N. M., "Face Veins based MCMT Technique for Personal Identification", International Journal of Intelligent Systems and Applications, Volume 7, Issue 9, pp. 57-72, 2015.
[57] Ray S, Mishra K N. and Dutta S, "Susceptible data classification and security reassurance in cloud-IoT based computing environment", Sadhna-Journal of Engineering Sciences Vol. 46, pp.1-24, 2021.

Authors’ Profiles

Soumya Ray is pursuing his Ph.D. in the Computer Science department from BIT Mesra, Ranchi. His research interest includes net security, big data security, cloud computing, fog computing, and the Internet of Things.

Dr. Kamta Nath Mishra
Computer Science & Engineering, Birla Institute of Technology, Jharkhand, India
Deoghar, Jharkhand, Pin 814142
Phone +91-9695052989

Dr. K. N. Mishra is a faculty member at B.I.T. mesra and has published more than forty research papers in journals and conferences of international repute. His research interest includes Biometric Systems, Image Processing, Analysis of Algo and Distributed Cloud Computing. Dr. Mishra is a professional member of IEEE Biometric Society USA, and ACM, USA.

Dr. Sandip Dutta is presently working as a professor and Head, Department of Computer Science &Engg, Birla Institute of Technology, Mesra, Ranchi. He has received his Ph.D. degree in the Computer Science Engg. from BIT Mesra, Ranchi. Prof. Dutta has more than 20 years of teaching experience. His research area includes Network Security, Cryptography, and Biometric Security. He has published more than forty research papers in reputed international journals, International Proceedings, and book chapters in his research area.

How to cite this paper: Soumya Ray, Kamta Nath Mishra, Sandip Dutta, "Sensitive Data Identification and Security Assurance in Cloud and IoT based Networks", International Journal of Computer Network and Information Security(IJCNIS), Vol.14, No.5, pp.11-27, 2022. DOI:10.5815/ijcnis.2022.05.02