Research Journal of Applied Sciences, Engineering and Technology 8(13): 1498-1509, 2014
DOI:10.19026/rjaset.8.1128
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2014 Maxwell Scientific Publication Corp.
Submitted: November 23, 2013 Accepted: March 08, 2014 Published: October 05, 2014

Research Article

Data Classification Based on Confidentiality in Virtual Cloud Environment

Munwar Ali Zardari, Low Tang Jung and Mohamed Nordin B. Zakaria
Department of CIS, Universiti Teknologi PETRONAS, Malaysia

Abstract: The aim of this study is to provide suitable security to data based on the security needs of data. It is very difficult to decide (in cloud) which data need what security and which data do not need security. However it will be easy to decide the security level for data after data classification according to their security level based on the characteristics of the data. In this study, we have proposed a data classification cloud model to solve data confidentiality issue in cloud computing environment. The data are classified into two major classes: sensitive and non-sensitive. The K-Nearest Neighbour (K-NN) classifier is used for data classification and the Rivest, Shamir and Adelman (RSA) algorithm is used to encrypt sensitive data. After implementing the proposed model, it is found that the confidentiality level of data is increased and this model is proved to be more cost and memory friendly for the users as well as for the cloud services providers. The data storage service is one of the cloud services where data servers are virtualized of all users. In a cloud server, the data are stored in two ways. First encrypt the received data and store on cloud servers. Second store data on the cloud servers without encryption. Both of these data storage methods can face data confidentiality issue, because the data have different values and characteristics that must be identified before sending to cloud severs.

Keywords: Cloud computing, data classification, data confidentiality/sensitivity, distributed computing, K-NN, non-sensitive, RSA

INTRODUCTION

Cloud Computing is an internet based distributed virtual environment. All computational operations are performed on cloud through the Internet (Rawat *et al.*, 2012). The cost of the resource management is more than the actual cost of the resources. So, it is often better to get the required resources by renting despite purchasing one's own resources. Basically, the cloud computing provides all IT resources for rent. The simple definition of cloud computing is: "A distributed virtual environment provides virtualization based IT-as-Services on rent". Beside all of the services like Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS), cloud also provides storage as a service, in which distributed database servers are available for rent to consumers. These services are available for all users without any business or data bias.

Consumers nowadays are using cloud services to avoid IT infrastructure purchasing and maintenance cost. A large amount of data can be stored on cloud. Cloud computing poses a number of challenging threats in the distributed storage model (Deepanchakaravarthi and Abburu, 2012). The data security is always the main challenging threat for quality of services and also stops the users to adopt cloud services (Rittinghouse and Ransome, 2009). In cloud storage, all kinds of data are stored on servers and they are stored through two storage methods. The first method is to encrypt received data and store on cloud servers. The second method is to store data on servers without encryption. These data storage methods can face data confidentiality issue. It is known that data are often not of the same type and have different properties and characteristics. In a cloud environment, a consumer's data are stored on remote servers that are not physically known by the consumer and there is high chance of confidentiality leakage. This study focuses on the confidentiality threat in the cloud environment. When a dataset is being transferred to cloud, it passes through a security mechanism, such as data encryption (without understanding the features of data) or directly being stored on servers without encryption. All data have different kinds of sensitivity levels. So, it would be non-technical to just send data into a cloud without understanding its security requirements. To address the security requirements of data, we have proposed a data classification model in the cloud environment to classify data according to its sensitivity level.

LITERATURE REVIEW

To show the importance of data security in cloud computing, the European Network and Information Security Agency (ENISA) published a report titled

Corresponding Author: Munwar Ali Zardari, Department of CIS, Universiti Teknologi PETRONAS, Malaysia

"Cloud Computing: Benefits, risks and recommendations for information security" in Nov-2009. In the report, ENISA found different cloud risks and their effects on the cloud consumers (Catteddu and Hogben, 2009). A crypto co-processor was suggested in (Ram and Sreenivaasan, 2010) to solve the data security threats in cloud. The crypto co-processor is a tool which provides security-as-a-service on demand controlled by a third party. Crypto co-processor allows the users to select the encryption technique to encrypt the data and divide data into different fixed chunks. This is to make hacker not knowing the starting and ending points of the data. But the limitation with this study is that the end user may not be technically savvy enough to select powerful technique for data encryption.

The single cloud is a central storage place for all consumers' data and the single central storage is easier to hack than compared to multiple storages. IBM proposed a new concept of inner-cloud in 2010. The inner-cloud is the clouds of a cloud. The inner-cloud storage model is more reliable and trustworthy as compared to a single cloud storage model (Cachin *et al*., 2010). In the inner-cloud model, the hash function and digital signature are hybridized to provide data authentication and integrity in a cloud. Whereas the data security key is divided and shared in multiple clouds; but this process of sharing of keys leads to a key issue when one cloud is not available. The integrated Data Protection as a Service (DPaaS) model is also used for data security (Song *et al*., 2012). DPaaS integrates information flow checking, encryption and application installation in cloud computing to avoid the implementation of the FDE and FHME techniques which are not affordable by small and medium enterprises and cloud service providers. The public cloud has still security challenges and data outsourcing is a big challenge. In data outsourcing, the user can not be sure about the location, data transaction accuracy and security of the stored data.

Most of these techniques i.e., discussed above work on data encryption for data security. To encrypt complete data, it is very expensive in the context of time and memory. It would be better to separate the sensitive data from the public data first and encrypt only the sensitive data.

The data classification is fundamental to risk assessment and useful for security control in organization (Etges and McNeil, 2006). Without understanding the importance of data, it is impossible to secure the business operation on data. Basically there are classes of data used by military. These classes are also known five-levels of data; these are unclassified, sensitive-but-unclassified, confidential, secret and top secret (Etges and McNeil, 2006). In July 2011, Machigan Technological University published a data classification policy "Data Classification and Handling Policy". In this report different data security controls are found for different data classes (Michigan Tech. University, 2011). All institutional data must be classified (according to its sensitivity, criticality and value) in to three classes that are confidential, internal/private and public (UTHSCSA, 2011; The California State University, 2011).

Classification of objects is an important area of research and of practical applications in a variety of fields, including pattern recognition and artificial intelligence, statistics, cognitive psychology, vision analysis and medicine (Keller *et al*., 1985; Hunt, 1975). There are numerous machine learning techniques that have been developed and investigated for classification. However, in many pattern recognition problems, the classification of an input pattern is based on the data where the respective sample size of each class is small. Moreover, the sample may possibly not be representative of the actual probability distribution, even if it is known (Keller *et al*., 1985). In such cases, there are many techniques that work on similarity and distance in feature space, for instance, clustering and discriminate analysis (Duba and Hart, 1973). In many areas, the K-Nearest Neighbour (K-NN) algorithm is used for classification of instances. K-NN is the simplest clustering technique with low complexity. This decision rule provides a simple nonparametric procedure for the assignment of a class label to the input pattern based on the class labels represented by the k-nearest neighbour of the vector. K-NN classification is more suitable for those problem domains characterized by data that is only partially exposed to the system prior to employment (Whitney and Dwyer III, 1966; Dasarathy, 1980).

PROBLEM STATEMENT

The cloud services are openly available for all kinds of organization. Different organizations and government of different countries store their sensitive data in cloud. In such scenarios data confidentiality is the most critical issue (Rawat *et al*., 2012; Wu *et al*., 2008). In cloud, data is shared and stored on centralized place, so it is easy for malicious users to access, delete and change the sensitive data. Data confidentiality threat increases in cloud, due to the increasing number of users (Hunt, 1975).

The literature reviewed seems not able to answer for the following question: How to classify data on the basis of confidentiality in cloud environment? It is very important to classify data to know that what data need to be secured and what data does not need any security such as public data.

CLOUDSIM SIMULATOR

CloudSim is a toolkit (library) for simulation of Cloud computing scenarios. It provides basic classes for describing data centres, virtual machines, applications, users, computational resources and policies for management of diverse parts of the system (e.g., scheduling and provisioning). These components can be put together for users to evaluate new strategies in utilization of Clouds (policies, scheduling algorithms, mapping and load balancing policies, etc.).

For simulation, we chose the CloudSim simulator integrated with an eclipse java development tool. We found CloudSim is the best simulator in the current simulation tools. CloudSim is a non-GUI tool and needs other supporting java tools like JDK6 and eclipse. Most of the researchers among the world are using the CloudSim simulator for cloud simulation.

PROPOSED MODEL

Machine learning techniques are mostly used in pattern recognition and data segmentation. In our proposed model we used the K-NN machine learning technique in the cloud computing environment to solve the data confidentiality problem.

To separate sensitive and non-sensitive data, the K-NN classifier is used in a designed simulation environment. The value of k is maintained to 1 for accuracy. After finding the sensitive and non-sensitive data, the sensitive data is further transferred to the RSA encryption algorithm for data encryption to protect sensitive data from unauthorized users. Therefore, the public data is directly allocated a Virtual Machine (VM) without encryption. The VM will process the data and communicate with storage servers for the data storage on the cloud servers. Most of the clouds implement data encryption techniques to protect data. But it is better to decide the security of the data based on the sensitive level of the data instead imposing encryption on complete data or just sending complete data into the server without any security. The classification technique in cloud will easily decide the security requirements of the data. In this way we can save our data from over-security and under-security situations and also save time and memory resources. In this study, we classified data into two different classes class 1 (non-sensitive data) and class 2 (sensitive data). Figure 1 shows the detailed steps to solve the data confidentiality issue in cloud computing. In our proposed model, we assume that the user is the one who verifies the results of the data classification that the model has classified data accurately. For the purpose of meeting security breach notification requirements, we followed the rules and parameters selection for confidential data defined by the "The California State University (2011)" in Patail and Behal (2012) and "University of Texas Health Science Center at San Antonio" in The California State University (2011).

Data classification: Data classification is a machine learning technique used to predict the class of the unclassified data. Data mining uses different tools to know the unknown, valid patterns and relationships in the dataset. These tools are mathematical algorithms, statistical models and Machine Learning (ML) algorithms (Phyu, 2009). Consequently, data mining consists of management, collection, prediction and analysis of the data. ML algorithms are described in to two classes: supervised and unsupervised.

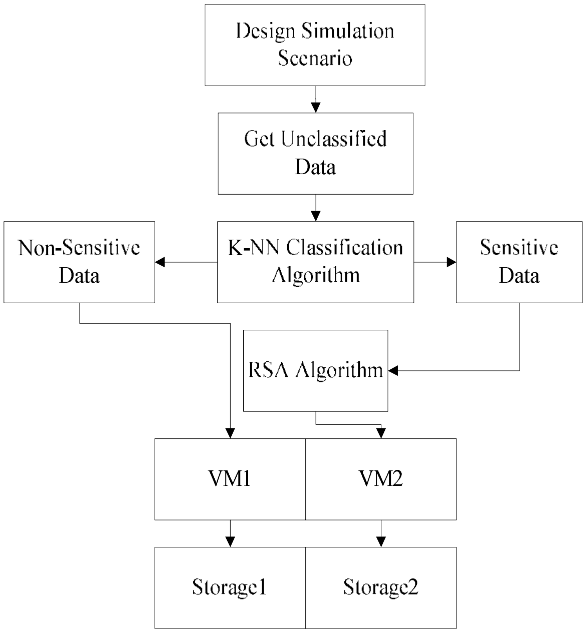

Fig. 1: Proposed model

Supervised learning: In supervised learning, classes are already defined. For supervised learning, first, a test dataset is defined which belongs to different classes. These classes are properly labelled with a specific name. Most of the data mining algorithms are supervised learning with a specific target variable. The supervised algorithm is given many values to compare similarity or find the distance between the test dataset and the input value, so it learns which input value belongs to which class.

Unsupervised learning: In unsupervised learning classes are not already defined but classification of the data is performed automatically. The unsupervised algorithm looks for similarity between two items in order to find whether they can be characterized as forming a group. These groups are called clusters. In simple words, in unsupervised learning, "no target variable is identified".

The classification of data in the context of confidentiality is the classification of data based on its sensitivity level and the impact to the organization that data be disclosed only authorized users. The data classification helps determine what baseline security requirements/controls are appropriate for safeguarding that data.

In this study, the data is classified into two classes, confidential and public (non-confidential) data. The classification of the data depends on the attributes of the data. The values of the sensitive attributes are classified as “sensitive” and values of the non-sensitive attributes are classified as “non-sensitive”. Michigan Technological University (2011) published a report titled “data classification and handling policy” in which they categorized data based on security into confidential data, internal/privacy data and public data (UTHSCSA, 2011). Another report published by University of Texas Health Science Centre at San Antonio (UTHSCSA, 2011) titled “protection by data classification security standard”, in which data is classified public, internal, confidential and confidential/high risk (The California State University, 2011). The California State University (2011) published a document titled “information security data classification standards”. This document describes the three classes of data regarding the data security placed on the particular types of information assets. These three classes are confidential, internal use and general (Patail and Behal, 2012). In this study we treat all data as confidential except public data because all other data need to be secure at different stages of data process with different methods. These protection methods can be secure area storage, lockable enclosure, and reasonable precautions to prevent access by non-staff and employees, password protection, encryption and so on.

Confidential data is a generalized term that typically represents data classified as confidential; according to the data classification scheme defined this document. This term is often interchangeably with sensitive data (UTHSCSA, 2011). Data should be classified confidential when discloser, alteration and destruction of that data cause a significant level of risk for the organization. Examples of confidential data include data protected by state or federal privacy regulations and data protected by confidentiality agreements. The highest level of security controls should be applied. The access of data must be granted only authorized persons or those persons affiliated to the organization that needs that data to perform their job. The access of confidential data must be controlled by the data owner and also provide access to desired users based on individual request.

The confidential data is highly sensitive data and may have different personal privacy considerations, or restricted by different laws i.e., state, federal, ISO, HIPAA and FERPA. Examples of confidential/sensitive data include official students’ grades and financial aid data, social security and credit card numbers and individuals’ health information.

The data must be classified as public data when unauthorized discloser, destruction and alteration of that data would result no risk to organization. Only supportive control is required to prevent unauthorized modification or destruction of public data. The public data is not treated as sensitive data, so it can be granted to any request or published without any security but integrity of public data must not be compromised. It must be protected from unauthorized integration. The examples of public data include directory information, course information and research publications.

For this simulation, the dataset titled “Adult. Data” is taken from (http://www.sgi.com/tech/mlc/db/, referred: Aug-2013). This dataset contains a company employees’ record. The dataset contains 5049 instances which are classified into two classes based on the data sensitivity level, for more details about data classification see the results. The data is classified based on the classification rules and data sensitive attributes discussed in UTHSCSA (2011), The California State University (2011), Patail and Behal (2012) and Etges and McNeil (2006); few of the rules are described in Table 1. Table 2 shows the

Table 1: Data security requirements

	Data security requirements	
Security control category	Public	Confidential
Sensitivity level	Low sensitive level	Highest sensitive level
Access controls	No restriction	Viewing and alteration restricted to authorized users Data owner permission required Authentication and authorized required to access Confidentiality agreement required
Network security	May reside on a public network	Protection with network firewall Protection with router ACLs Servers hosting the data cannot be visible to the entire internet, nor to unprotected subnets like the residence halls and guest wireless networks
Virtual environment	No need any security except integrity	Can not share the same virtual host environment with guest virtual servers of other security classification
Data storage	Store on virtual cloud server Prevent from data loss	Storage on secure virtual cloud server required Encryption is required Prevent from data loss Keep data backup
Transmission	No restrictions	Encryption required: for examples, via SSL or secure file transaction protocols
Backup	Backup required	Daily backup required Off-site storage in a secure location required in cloud

Table 2: Dataset attributes

Data set attributes	Class
Age	Sensitive
Work class	Sensitive
Final weight	Sensitive
Education	Non-sensitive
Education-number	Sensitive
Marital-status	Sensitive
Occupation	Non-sensitive
Relationship	Sensitive
Race	Non-sensitive
Sex	Non-sensitive
Capital-gain	Sensitive
Capital-loss	Sensitive
Hours-per-week	Sensitive
Native-country	Non-sensitive
Salary	Sensitive

Table 3: Dataset details

Dataset	Instances	Categorical features	Numerical features	Missing values
Adult	5049	32	6	761

Table 4: Classification rules/conditions

Attribute 1	Attribute 2	Class
Age: x years	Sex: female	Sensitive
Sex: female	Marital-status: x	Sensitive
Capital-gain	Capital-loss	Sensitive
Capital-gain	-	Sensitive
Capital-loss	-	Sensitive
Age	Work-class	Non-sensitive
Education	Native-country	Non-sensitive
Race	-	Non-sensitive

“Adult.data” dataset parameters. These parameters are labelled as sensitive and non-sensitive based on predefined rule defined in The California State University (2011) and Patail and Behal (2012).

Table 3 shows the information of dataset. The dataset contains following attributes with different values.

Table 4 shows the examples of sensitive and non-sensitive data. The proposed model classifies data into sensitive and non-sensitive by following the data security conditions/rules given in Table 4. For example:

> If sex = female and marital-status = x then classify it sensitive.
>
> If capital-gain (and) / (or) capital-loss is there then classify it as sensitive.
>
> If age and work-class is there then classify it as non-sensitive.
>
> If education and native-country is given classify as non-sensitive.

Data security requirements: Data can be sensitive (confidential) or non-sensitive (public) with different security requirements. These security requirements depend on the sensitivity levels of data. The sensitive data need high level of security whereas the public data do not need any security except integrity protection. It is better to understand the exact security requirements of data before transferring it to the cloud which is only possible through machine learning techniques.

For data classification, several data handling requirements are defined to appropriate safeguard the information (UTHSCSA, 2011). It is important to understand overall sensitivity data encompasses based on confidentiality. The Table 1 determines safeguard requirements for data protection based on data classification.

K-NN classifier: The K-Nearest Neighbour (K-NN) is a supervised machine learning technique; it is widely used for classification, pattern recognition, prediction and estimation. K-NN totally depends on instance-based learning, where a set of training data is stored for the classification of new unclassified datasets (Larose, 2005). It is the simplest iterative technique to classify unclassified datasets into user specified classes, k. This algorithm discovered by several researchers across different disciplines, most notably, Forgey (1965), Friedman and Rubin (1967), McQueen (1967) and Lloyd (1957).

The K-NN algorithm has a set of n labelled samples; *n* is the number of data items in the set. This can be represented as:

$$D = \{d_1, d_2, d_3, ..., d_n\}$$

where, *D* is the set of total samples and d_1, d_2, d_3,…, d_n are different samples. The *D* must be assigned n labels. The set of n labelled samples can be represented as:

$$D = \{d_1, d_2, d_3, ..., d_3 \mid c_1, c_2, c_3, ... c_n\}$$

where, c_1, c_2, c_3,…, c_n are different classes for the targeted values.

How the K-NN algorithm works:

Step 1: Determine the set of *n* labelled samples: D
Step 2: Determine value of K
Step 3: Calculate the distance between the new input and all of the training data
Step 4: Sort the distance and determine the K-nearest neighbours based on the K-th minimum distance
Step 5: Find the classes of those neighbours
Step 6: Determine the class of the new input based on a majority vote

Explanation of steps:

Step 1: First, define an array-list with number of attributes of training dataset. These attribute also known as samples. It is important at this step to define the classes to these attributes because it will be easy to determine the class of input value through finding similarity with attributes (samples). The sample class will be assigned to targeted value if they are same or having minimum distance.

Step 2: After determining samples and classes, define the value of *K*. Here *K* is the number of nearest neighbours to an input value (new value need to be classified). The value of *K* can be any value but select minimum value for more accuracy.

Step 3: Calculate the distance or similarity of training data attributes and a new input value to determine the class of the new input value. For this simulation, we used Euclidean's distance formula to calculate distance. With distance we also measured the similarity between training data and new input value for more accuracy. The Euclidean distance can be calculated as:

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where, $D(x, y)$ is distance between x and y and $x = x_1, x_2, x_3, ..., x_n$, $y = y_1, y_2, y_3, ..., y_n$ show the n attribute values of two records x and y. In other words:

x_i = Number of labelled samples.
y_i = Number of new input attributes/values need to be classify.

Step 4: During distance calculation, calculate the distance of new input value with all samples one by one and store every result of distance in an array. After that, sort the array in ascending order and find specified K nearest neighbours (the samples which have minimum distance with new input value). The class of new input value will be decided based on this step.

Step 5: Get the classes of K-nearest neighbours from set of classes already defined with *D* samples based on minimum distances.

Step 6: After getting classes of nearest neighbours of new input value, find the classes of majority votes in the nearest neighbours. The decision of class for new input value will be taken based on majority votes. The class which has vote in majority will be assigned to new input value.

After classifying data using K-NN algorithm encrypt sensitive data using RSA algorithm discussed in section RSA algorithm. For further process we used following steps to process data in cloud.
Steps:

- Assign encrypted data and unencrypted data to cloudlets as:
  - $cloudlet\_1 \leftarrow encrypted - data\, and$
  - $cloudlet\_2 \leftarrow unencrypted - data$
- Select the VMs which are free or have still enough capacity to handle cloudlets:
  - If (VM is free and has enough capacity to handle cloudlet):
    - Assign cloudlet_1 to VM_1: $VM\_1 \leftarrow Cloudlet\_1$
    - Assign cloudlet_2 to VM_2: $VM\_2 \leftarrow Cloudlet\_2$
- Choose data center (s)
- Assign VMs to selected data center (s) through broker
- Store data on data center (s)

In K-NN algorithm K is the number of neighours closest to the new record. The value of k can be any positive integer number. The accuracy of the algorithm depends on the value of K, the change in the value of K directly affect the accuracy in the selection of the class. In K-NN algorithm it is difficult to select an appropriate value for K. For example: If k = 3, it will find three nearest neighbours to the input value and assign the class to the input value on the basis of the majority votes among the three votes. In another case, if K = 5, the algorithm will find 5 nearest neighbours to the input value and assign the class to the input value on the basis of the majority vote among the five closest values' class. In this case, let us suppose that two items belong to one class and the remaining 3 items belong to other classes; according to K-NN, the new input value will be assigned to the majority vote but the input value may belong to the minority vote. This is a big disadvantage of the K-NN algorithm. The same scenario is described in Fig. 1. The red colour plus (+) symbol is considered unclassified data in both the "a" and "b" figures. In figure "a", we suppose k = 3. After classification the red plus (+) belongs to class PLUS, because it's two neighbours belong to the PLUS class. Whereas, in figure "b" we suppose that the value of k = 7. After classification, the red plus (+) belongs to the MINUS class because, it's four neighbours (majority) belongs to the MINUS class; but in all actuality it should be assigned to the PLUS class (Fig. 2).

RSA algorithm: Data confidentiality is grown and to become a non-resolvable threat in cloud. Data confidentiality threat appears in complete life cycle of data transaction in cloud. Data security is a technique used to keep data secure from internal and external threats in cloud and also in transaction life cycle. RSA is an asymmetric encryption algorithm to encrypt data with the help of public key. RSA is developed by three scientists, Rom Rivert, Adi Shamir and Leonard in 1978 (Patail and Behal, 2012). The public key is used to encrypt data and it can be known publically. The private key a secret key only know to receiver and used to decrypt data. The RSA algorithm is based on complicated factoring of large two prime numbers p and q. It is mandatory to choose large prime numbers

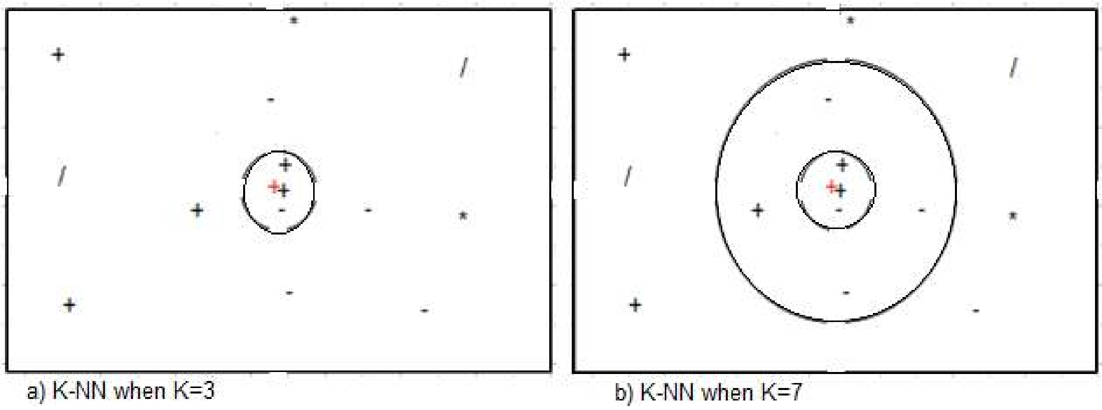

Fig. 2: K-NN classifier with different values of K

because the factors of large prime numbers are very difficult to break. So, the large prime number the high security is directly proportions to each other in this situation. The private and public key generation is also dependent on prime numbers. The concept of factorization of two prime numbers makes RSA more secure.

How RSA algorithm works:

- Choose two large prime numbers p and q
- Compute n = p.q
- Compute Ø (n) = (p-1) (q-1)
- Choose e such that 1<e<Ø (n) and e and n are coprime
- gcd (e, Ø (n) = 1)
- Compute a value fo d such that e.d = 1 mod Ø (n) & 0<d<n
- Public key (e. n)
- Private key (d, n)
- The Encryption of plain text c = m^e mod n
- Decryption d = c^d mod n

CLOUD SIMULATION ENVIRONMENT

The Cloud Sim simulator was used for simulation purposes. Figure 3 shows the proposed simulation environment for cloud service providers to solve data sensitivity/confidentiality issue in cloud computing. At the bottom, we used a Cloud Sim engine to run the simulation. The Virtual Machine Manager (VMM) was used to manage and allocate VMs to cloudlets (cloud tasks). The number of cloud items used for the simulation is given in Table 5.

Table 5: Cloud items and quantity

Items	Quantity
Data centre	2
Host	1
Broker	1
VM	2
VMM (Xen)	1
Cloudlet	2

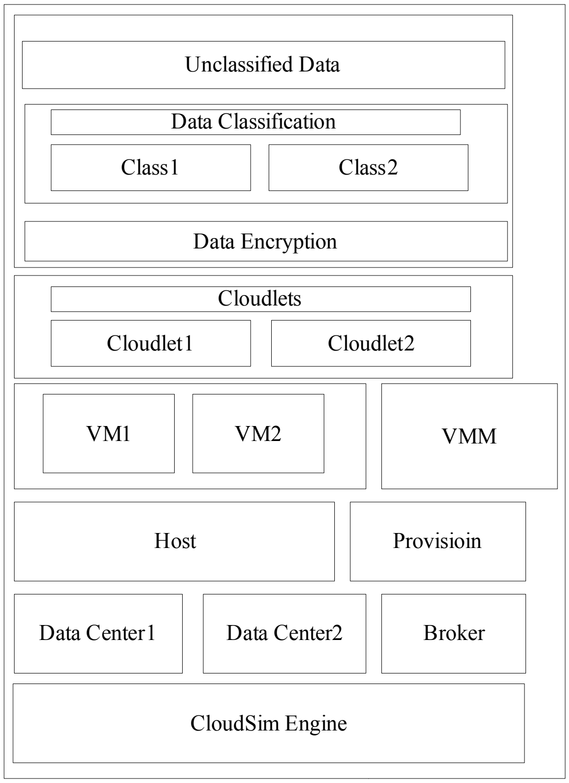

Fig. 3: Simulation environment

Data centre: This can handle many hosts in terms of managing VMs in their life cycles. In Cloud Sim, the host represents a physical computing server which is assigned MIPS (Millions of Instructions per Second), storage, memory and an allocation of processing cores. The host supports single core and multi-core simulation. In single core simulation we specify a single processor and in multi-core simulation we have more than one processor.

Table 6: SaaS properties

Cloudlet ID	Length (in bytes)	Input file size (in bytes)	Output file size (in bytes)
0	4000	158	158
1	3000	139	139

Table 7: PaaS properties for virtualization management

VM ID	MIPS	Image size (MB)	Bandwidth (Mbps)	Pes No.	VMM
0	100	10000	1000	1	Xen
1	100	10000	1000	1	Xen

Table 8: IaaS properties for cloud simulation

DC ID	RAM (Mb)	Storage	Data architecture	OS	Bandwidth
2	2048	10000000	X86	Linux	10000
3	2048	10000000	X86	Linux	10000

Virtual machine: One of the main aspects that differentiate cloud from other distributed environments is the virtual machine, which creates a virtual environment. All requests are modelled by a simulator in the data centres. These requests are assigned through the VM to the data centres. The VM has a complete lifecycle in the simulator as well as in real cloud environment. The VM lifecycle consists of: provisioning of host to a VM, provisioning of cloudlets to VM, VM creation, VM destruction and VM migration. The VM allocation to the application specific host is the responsibility of the VM provisioner component.

Broker: This acts on the behalf of the consumers to sift through various offerings and decide on what are the best for the consumer. In simple words, it is a third party which creates an understanding environment between the consumer and the CSP.

Provision: This allocates VMs to the application specific host.

Cloud let: This is a task unit sent to the VM to process and allocate to a specific host.

Cloud service properties and description: Before simulation it is important to set the properties of all three service models Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). Table 6 shows the properties of the SaaS modeller which was deployed on a VM in Cloud Sim. In the SaaS modeller every cloudlet has a specific identification for the VM. Here "ID" represents to a specific cloudlet and length is the size of the cloudlet. The size of the input and output file is given in Bytes.

Length: The length or size (in MI) of this cloudlet is to be executed in a data centre.

Input file size: The file size (in Bytes) of the current cloudlet BEFORE being submitted to a data centre.

Output file size: This is the file size (in Bytes) of the current cloudlet AFTER execution.

Table 7 shows the properties of PaaS for the application deployment which contains VM properties. It shows the processing power of the physical computing node which is assigned at the virtual machine level, VM image size (in MB), amount of bandwidth and the number of cores in which the MIPS' power is shared at the VM level to run the cloudlet. The VMs are managed by VMM.

Machine Instructions Per Second (MIPS): This is the processing power assigned to the VM to execute the instructions according to the specified MIPS.

Image size: The Image Size (in MB) is the VM image size that is represented to a virtual hard disk file that is used as a template for creating a VM.

Pes number: The Pes number is the number of processors used at the VM level.

It is also important to use better and stronger infrastructure resources in cloud for better computation and response time. The available resources at this level put a limit on the SaaS modeller requirement, i.e., resources allocated at the VM level can't exceed this limit. Table 8 shows the IaaS properties and their values, where "DC ID" is the data centre identity which is assigned to the VM.

RESULTS AND DISCUSSION

In this section, we discuss the results obtained after the implementation of two algorithms to improve and manage data confidentiality in a cloud environment. The data selected for this study is the employees' records of an organisation. This data was taken from (http://www.sgi.com/tech/mlc/db/, referred: Aug-2013), which contains different types of datasets mostly used by the research community. Table 9 shows the details of the file before and after classification and also the accuracy of algorithm. The accuracy of 1-NN is 78.8007 and 21.1923% instances are incorrectly classified. Here we used 1-NN which shows that that value of k for this data is 1. For algorithm performance accuracy we used 10-fold cross validation to measure the percentage correctly and incorrectly classified instances of dataset. The total size of the file was 512 KB and total instances in file were 5094. The K-NN classifier was used to classify the data into two classes: sensitive and non-sensitive. After classification, the non-sensitive data was labelled as "Class1" and the sensitive data was labelled as "Class2". The time taken by the K-NN classifier was 1075 msec as shown in Table 6. The results have been categorised into 3 cases which are discussed below.

Case 1:

Simple classification: This case only defines the simple classification of the data and simple simulation of these classes.

Table 9: Classification of data

		After classification						
Before classification		Class 1		Class 2				
Total size of file (KB)	Total number of instances	Size (KB)	Instances	Size (KB)	Instances	K-NN time (msec)	Correctly classified instances	Incorrect classified instances
512	5049	352	3405	160	1644	1075	78.8077%	21.1923%

Table 10: Cloud simulation

Cloudlet ID	VM ID	Data centre ID	Status	Start time (msec)	Finish time (msec)	Total time (msec)	Total time (taken by both cloudlets)
0	0	2	Success	0	3400	3400	5040 msec
1	1	3	Success	0	1640	1640	

Table 11: Total simulation time

Classification time (msec)	Total time taken by cloudlets after classification (msec)	Total simulation time (msec)
1075	5040	6115

Table 12: Simulation of encrypted sensitive data

Cloudlet ID	Status	Data centre ID	VM ID	Start time (msec)	Finish time (msec)	Total time (msec)	Total time (taken by both cloudlets) (msec)
0	Success	2	0	0	3400	3400	170400
1	Success	3	1	0	167000	167000	

Table 10 shows the results of the simulation. Here, cloudlet ID = 0 represents the Class 1 data and cloudlet ID = 1 represents the Class 2 data; these cloudlets are assigned data of different classes as shown below:

$$(cloudlet\_ID = 0) \leftarrow class1\_data$$

$$(cloudlet\_ID = 1) \leftarrow class2\_data$$

If the status of the simulation was "SUCCESS", it means that the simulation was performed successfully. Each class data was assigned to a different VM and data centre.

Start time: This is the time at which the VM starts processing the received cloudlet. This time can change for each cloudlet when multiple cloudlets were assigned to a single VM. But here, the start time was zero (0) because the VM was not busy before with any other cloudlets.

Finish time: This is the time that the cloudlet finished being processed by the VM. The finish time depended on the size of the cloudlet.

Total time: This is the total time slot taken by a cloudlet whilst being processed. The total time was calculated from Eq. (1):

$$Time\_C_i = FtC_i - StC_i \tag{1}$$

$$TotalTime\_C_{1,2} = Time\_C_1 + Time\_C_2 \tag{2}$$

The Eq. (2) can be written as:

$$TotalTime\_C_{1,2} = \sum_{i=1}^{2} Time\_C_i \tag{3}$$

where,

$Time\_C_i$ = The total time taken by cloudlets
FtC_i = The finish time
StC_i = The starting time

Equation (2) was used to calculate the total time taken by both cloudlets during simulation. Where, TotalTime_$C_{1,2}$ is the total time that was taken by both cloudlets, Time_C_1 is the total time that was taken by cloudlet1 and Time_C_2 is the total time that was taken cloudlet 2. In Eq. (1), (2) and (3), i = 1, 2.

The initial start time of both cloudlets was zero (0) because these are the first cloudlets to VMs. In simple: start time is the starting time of a VM to process a cloudlet. The total simulation time has been calculated in Table 11. The total simulation time is the sum of the total classification time and the total time taken by both cloudlets (both classes) after classification. This calculation was performed by using Eq. (4):

$$TST = CT + TC_i \tag{4}$$

where,

TST : The total time that was taken by simulation
CT : The total time that was taken by classification
TC_i : The time that was taken by the cloudlets after classification

Case 2:

Encryption of class 2 (sensitive data): In this case, after the classification the sensitive data (class 2) was encrypted using the RSA algorithm and the time and size of Class 2 were compared with the original size and time. Table 12 shows the simulation results of class 1 and 2. Class 1 data took the same time as it has

Table 13: RSA calculation

Class 2 size (KB)	After encryption class 2 size (KB)	After encryption: 16335 KB was considered to the size of the number of normal instances?	Encryption time (msec)
160	16335	167883 instances	873824

Table 14: Total simulation time

Classification time (msec)	Time taken by RSA (msec)	Time taken by cloudlets after classification and encryption (msec)	Total simulation time (msec)
1075	873824	170400	1045299

Table 15: Data encryption without classification

Total data size (KB)	Total time taken by RSA (msec)	Total data size after encryption (KB)
512	2796237	50166.86 This size was considered to be the size of 5155810 instances

Table 16: Total time taken by data

Time taken by RSA (msec)	Total time taken by cloudlet (msec)	Total simulation time (msec)
2796237	5155800	7952037

taken in case 1 but the simulation time of class 2 data changed after the encryption. Equation (3) was used to calculate the total simulation time taken by both cloudlets.

After encryption, the size of the class 2 data (sensitive data) changed as shown in Table 13. The original size (before encryption) of the data was 160 KB. The 160 KB was the size of 1644 instances as shown in Table 6. After encryption, the size of the data increased up to 16335 KB, which was assumed the size of 167883 normal instances (non-encrypted instances). The total time taken by the RSA to encrypt 160 KB was 873824 msec.

Table 14 shows the total time of the simulation with classification and encryption time. The total simulation time was calculated using the Eq. (5):

$$TST = \left(CT + ET + TC_i\right) \tag{5}$$

where,

TST : The total simulation time
CT : The classification time
ET : The encryption time
TC_i : The time that was taken by both cloudlets after classification and encryption

Case 3:

Encryption of all data (without classification): In case 2, we explained the complete simulation process with the encryption of only the sensitive data, where we proposed the data classification technique to classify the data on the basis of the sensitivity level. In this case, we followed a traditional way of cloud computing in which the complete data (public and private) was encrypted without identifying the security requirements of the data.

Table 15 shows the encryption details of the total 512 KB of data. The complete data of the consumer was encrypted by the CSP and stored in the cloud server. This situation led to many issues like the extra consumption of resources. After the encryption of the data, the size of the data was increased from 512 to 50166.86 KB as shown in Table 15.

Equation (6) was used to calculate the total simulation time in this case. The details of the simulation timing are shown Table 16. The simulation time and the size of the data increased after the encryption.

The total simulation time in case 2 was 1045299 msec as shown in Table 14 but if we encrypt the whole data (ignore the classification) the simulation time was increased up to 7952037 msec as shown in Table 15:

$$TST = ET + TC_i \tag{6}$$

where,

TST = Total simulation time
ET = Encryption time
TC_i = Time taken by cloudlets after encryption or classification

Comparison of cases: After the discussion of all three (3) cases, Fig. 4 and 5 show the comparison of these cases based on the data size (in KB) and cloud processing time (in milliseconds).

Figure 4 shows the comparison of cases on the bases of data size. The data size in case1 was 512 KB (but no security for any class). Whereas, in case 2 the sensitive data was encrypted after its identification (using classification) and the size of the encrypted data had increased. So, the total size of the stored data was also increased up to 16687 KB as shown in Fig. 4. In case 3, without identifying the security requirements of the data, the complete data is encrypted. After encryption, the size of the stored data was increased. The case 3 and 1 (without classification) are the traditional cases in cloud which are mostly followed by cloud vendors. Figure 5 shows the comparison of the cases based on the simulation time. The total simulation time that was taken in case1 was 6115 msec; in case 2, the total simulation time that was taken by the data was 1045299 msec and in case 3, the total simulation time that was taken by the data was 7952037 msec. The simulation time is directly proportional to the size of the data which can be written as:

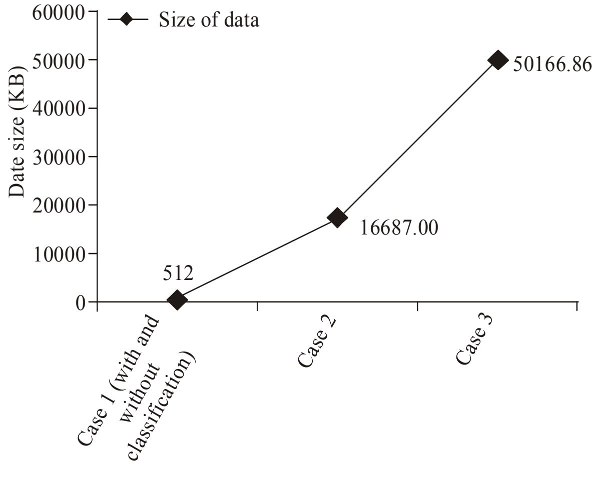

Fig. 4: Comparison of cases on the basis of data size

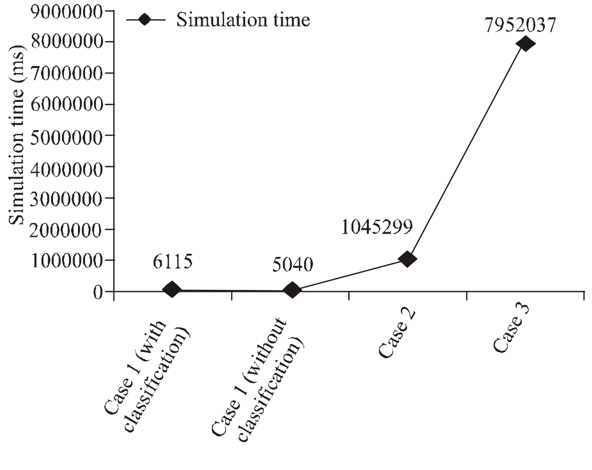

Fig. 5: Comparison of cases on the basis of simulation time

$$simulation\_time \propto data\_size$$

After the comparison, the proposed case 2 was better than case 1 and 3. In case 2, after the classification, we encrypted only the data which required security. The other data was considered as public. But, in case 1 all of the data was transferred into cloud without taking into consideration the sensitivity of the data. In case 1, the sensitive data was not secure because the sensitive data was treated as public data. Whereas, in class 3, the complete data was considered sensitive as it consumed more resources as shown in Fig. 4 and 5.

We proved that data classification in cloud is very important in order to know which data need security and which data do not need security. After applying the classification, we were able to handle the following situations easily:

- **The cloud services consumer may under value the data:** This situation takes place when a consumer does not know about sensitive and non-sensitive data; he/she just sends data into the cloud without any encryption. The sensitive data is also transferred into the cloud along with the non-sensitive data.
- **The cloud services consumer may over value the data:** This situation takes place when a consumer sends data into the cloud and applies the data security service (encryption) on the whole data. Here, the consumer encrypts all of his/her data even the public data which do not need any security.

Advantages of the proposed model in cloud:

- To know the sensitivity level of the data
- To select security according to the sensitivity level of the data
- To avoid the massive use of an expensive data encryption technique
- To avoid over-fit data security for data being stored in cloud servers
- To avoid under-fit data security for data being stored in cloud servers
- To save memory, time and extra payment resources

We proposed a new model with data classification in cloud, which classifies data based on the sensitivity level of data. The results of the proposed model are more favourable as compared to the ordinary methods of data storage in the cloud. For data classification, we proposed a new layer of classification in cloud virtual environment. In future, this layer can be used as a new cloud service i.e., Data-classification-as-a-Service (DcaaS) for cloud users.

CONCLUSION

In this study, we have proposed a resource and cost friendly model for cloud computing with data classification service. The focus of this study was to improve the confidentiality level of data during data outsourcing. The basic contribution of this model is confidentiality based data confidentiality with machine learning technique (data classification). For data classification, the K-NN classifier is used to classify data based on the security requirements of the data. The data is classified in to two classes: sensitive and non-sensitive data. The sensitive class is encrypted using the RSA algorithm whereas the non-sensitive data is directly stored on the cloud servers.

We analyzed that the proposed case 2 is best case among three cases. Case 2 is developed following our proposed model. The best way to improve the confidentially and the computational capacity of the VMs and servers is through data classification. In this way, it is easy to know which data need what security and which data do not need any security. Without data classification, the consumer may over-secure or under-secure his/her data. The proposed model has been implemented in a designed simulation environment using a CloudSim simulator. The simulation results were found to be strongly supporting our proposed model and confidentiality based data classification. Furthermore, to our best knowledge this proposed model is the first model in cloud computing with a data classification technique to improve the security of data. In future work, we will implement our own machine learning algorithm with multiple datasets in cloud environment and compare their performance and accuracy based on our proposed model.

REFERENCES

Cachin, C., R. Haas and M. Vukolic, 2010. Dependable Storage in the Intercloud. IBM Research Report RZ 3783.

Catteddu, D. and G. Hogben, 2009. Cloud Computing: Benefits, Risks and Recommendations for Information Security, ENISA, Crete, pp: 1-125.

Dasarathy, B.V., 1980. Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments. IEEE T. Pattern Anal., PAMI-2(1): 67-71.

Deepanchakaravarthi, P. and S. Abburu, 2012. An approach for data storage security in cloud computing. IJCSI Int. J. Comput. Sci. Issues, 9(2): 1694-0814.

Duba, R.O. and P.E Hart, 1973. Pattern Classification and Scene Analysis. John Wily and Sons Inc., New York.

Etges, R. and K. McNeil, 2006. Understanding data classification based on business and security requirements. J. Online, 5: 1-8.

Forgey, E., 1965. Cluster analysis of multivariate data: Efficiency vs. Interpretability of classification. Biometrics, 21: 768, 1965.

Friedman, H.P. and J. Rubin, 1967. On some invariant criteria for grouping data. J. Am. Stat. Assoc., 62: 1159-1178.

Hunt, E., 1975. Artificial Intelligence. Academic Press, New York.

Keller, J.M., M.R. Gray and J.A. Givens, 1985. A fussy-K-nearest neighbor algorithm. IEEE T. Syst. Man Cyb., SMC-15(4): 580-585.

Larose, D.T., 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley and Sons, Inc., Hoboken, New Jersey, pp: 90-106.

Lloyd, S.P., 1957. Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meeting Atlantic City, NJ, September 1957. Also, IEEE T. Inform Theory (Special Issue on Quantization), IT-28: 129-137.

McQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceeding of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1: 281-296.

Michigan Technological University, 2011. Information Technology Services and Security, 2011. Data Classification and Handling Policy.

Patail, A. and S. Behal, 2012. RSA Algorithm achievement with federal information processing signature for data protection in cloud computing. Int. J. Comput. Technol., 3: 34-38.

Phyu, T.N., 2009. Survey of classification techniques in data mining. Proceeding of the International MultiConference of Engineers and Computer Scientists. Hong Kong, Vol. 1.

Ram, C.P. and G. Sreenivaasan, 2010. Security as a service (SasS): Securing user data by coprocessor and distributing the data. Proceeding of Trendz in Information Sciences and Computing (TISC, 2010), pp: 152-155.

Rawat, P.S., G.P. Saroha and V. Barthwal, 2012. Quality of service evaluation of Saas modeler (Cloudlet) running on virtual cloud computing environment using CloudSim. Int. J. Comput. Appl., 53(13): 35-38.

Rittinghouse, J.W. and J.F. Ransome, 2009. Cloud Computing Implementation, Management, Security. CRC Press by Taylor and Francis Group, LLC.

Song, D., E. Shi, I. Fischer and U. Shankar, 2012. Cloud data protection for the masses. IEEE Comput. Soc., 45(1): 39-45.

The California State University, 2011. Information Security Data Classification.

UTHSCSA, 2011. Protection by data classification security Standard. Data Classification Report.

Whitney, A. and S.J. Dwyer III, 1966. Performance and implementation of K-nearest neighbor decision rule with incorrectly identified training samples. Proceeding of 4th Allerton Conference Circuits Band System Theory.

Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand and D. Steinberg, 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst., 14: 1-37.