Survey on Data Classification and Data Encryption Techniques Used in Cloud Computing

Prakash Sawle
ME Student
Department of IT
MIT, Pune

Trupti Baraskar
Assistant Professor
Department of IT
MIT, Pune

ABSTRACT
Cloud computing is an imminent revolution in information technology (IT) industry because of its performance, accessibility, low cost and many other luxury features. Security of data in the cloud is one of the major issues which acts as barrier in the implementation of cloud computing. In past years, a number of research works have targeted this problem. In this paper discuss some of the data classification techniques widely used in cloud computing. The objective of data classification is to find out the required level of security for data and to protect data by providing sufficient level of security according to the risk levels of data. In this paper also discuss a survey of existing solutions for security problem, discuss their advantages, and point out any disadvantages for future research. Specifically, focus on the use of encryption techniques, and provide a comparative study of the major encryption techniques.

Keywords
Cloud Computing, Data Classification, Data confidentiality, Cryptography, Caesar Cipher, Vigenere Cipher, fully Homomorphic, and Hierarchical Identity Based Encryption (HIBE).

1. INTRODUCTION

1.1 Cloud Computing
Cloud primarily refers to the saving of customers' data to an off-site storage system that is preserved by a third party. This means instead of storing data on customer computer's hard disk or other storage devices, client save it to a remote server storage system where the internet provides the connection between the user's computer and the remote server storage system. Computers in the cloud are constructed to work simultaneously and the various applications use the collective computing power as if they are running on a cloud using the concept of virtualization. In this model customer to the cloud to access information technology resources which are priced and provided on-demand. Cloud computing services such as Amazon EC2 and Google App Engine are built to take advantage of the already existing infrastructure of their respective company [1].

According to the official NIST definition, "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [2].

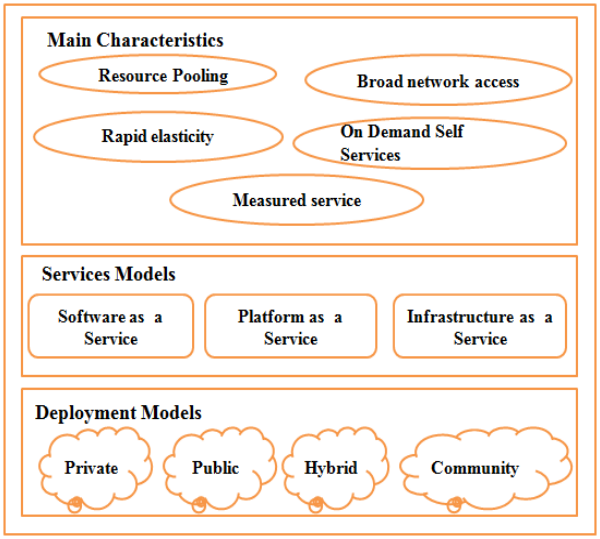

Figure1: Model of NIST definition of cloud computing

1.1.1 Main Characteristics

On-demand self-service
Computing capabilities provision made by consumer unilaterally, such as server time and network storage, as required automatically without requiring human dealings with each service provider.

Broad network access
The heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations) accessed through standard mechanisms of capabilities is available over the network.

Resource pooling
The service provider's computing resources are pooled to give out multiple consumers, with dynamically assigned and reassigned to different physical and virtual resources according to consumer demand. Examples of resources include storage, processing, memory, and network bandwidth.

Rapid Elasticity
The users can rapidly increase and decrease their computing resources as per needed.

Measured Service
The cloud systems control and optimize resource automatically using by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g. processing, storage, bandwidth, and active user accounts). The cloud systems also monitored controlled of resource usages and reported, providing transparency for both the provider and consumer of the utilized service.

1.1.2 Service Models

Software as a Service (SaaS)
This service provides the software to the consumer. Consumer doesn't require installing the software on their own machines and they can use the software directly from the cloud using internet. Examples Google Apps, Zoho, Salesforce.com,.

Cloud Platform as a Service (PaaS)
This provides the platform to the clients so that they can make their own applications on this platform and also provide rent server for applications. Examples Google App Engine, Windows Azure, Aptana Cloud.

Cloud Infrastructure as a Service (IaaS)
This service provides to user to rent the infrastructures like the storage system and computation resources, cloud provider examples Dropbox, Amazon Web Services, Mozy, Akamai.

1.1.3 Deployment Models

Public clouds
Service providers are offered for general public over the internet. Like Amazon, IBM's Blue Cloud, Google App Engine, Windows Azure etc. The infrastructure available can use by general public via internet.

Private clouds
These clouds are accessed by particular organization which are managed by third party or internally and hosted externally or internally like amazon (EC2).

Hybrid clouds
These clouds are a combination of private and public cloud. In these clouds critical data stored in private clouds and applications is hosted in public clouds.

Community cloud
It is established for particular community it can be Hybrid, private or public clouds [2].

Security is one of the biggest challenges faced in cloud computing. The security of data is a major issue which acts as an impediment in implementation of cloud computing. When we are storing data in the cloud some time critical data having high security and non-critical data need low security. To overcome this problem used data classification methods based on the cryptographic parameters.

This paper is structured as follows: Section2 data classification and survey of data classification techniques used in cloud computing.Section3 survey of existing data encryption techniques used in the cloud computing .Section4 comparison existing data encryption techniques.Section5 conclusion of this paper.

2. DATA CLASSIFICATIONS TECHNIQUES

2.1 Data Classification

The objective of data classification is to establish the required level of security for data and to protect data by providing a sufficient level of security according to the risk levels of data. Classification of data helps in defining the baseline security controls for protecting the data. The organization's information system must be carefully scrutinized and classified based on its level of sensitivity and the effectiveness of the organization if the data are disclosed, modified or destroyed without authorization. Classification identifies and splits the most sensitive data from less sensitive data [1].

2.2 Different Data Classification Techniques

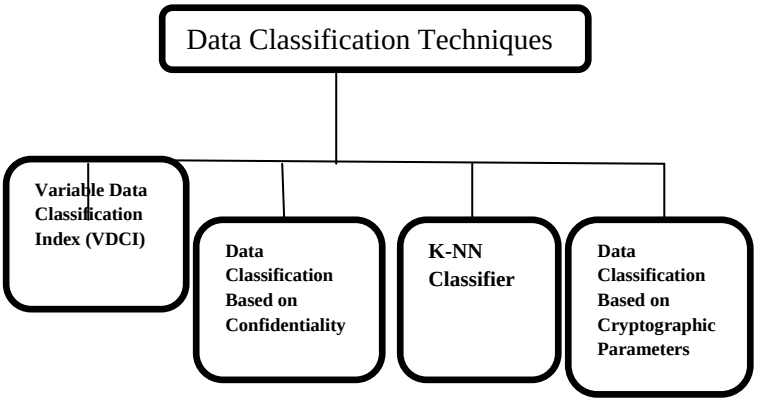

Figure 2 .Different Data Classifications Techniques

2.2.1 Variable Data Classification Index (VDCI)

Faraz Fatemi Moghaddam et al. proposed the Variable Data Classification Index (VDCI) is a variable data based on main three parameters and its own several sub-parameters. This value is calculated using the history of stored data instead of data owner or system administrator. Using determine the values of all three cryptographic parameters (i.e. availability, integrity and confidentiality).

Confidentiality	**Availability**	**Integrity**
CT: Number of users that are accessed to the cloud. CA: Number of users that are accessed to a specified file (FILE). CE: Number of users that can edit FILE. CR: Number of revoked users from accessing to FILE.	AD: The total number of downloading FILE. AU: The total number of re-uploading FILE. AT: A defined number by system administrator or data owner which provides the maximum number of download or upload processes.	IT: Number of requests to access FILE. IA: Number of un-authorized requests to access FILE.
C = [(CA+CT)/CT] + [((2×CE) +CT) /CT] + [CT/(CR+CT)]	A = [((2×AU) + AT) /AT] + [(AD+AT)/AT]	I = (IU+ (2×IA))/IA
The range of C is between $3 < C \leq 6$	The range of A is between $2 < A \leq 5$	The range of I is between $2 < A \leq 3$
VDCI = C + (3×A)/2 – (1/I)×10		

The value of VDCI was calculated concerning the inverse proportional of Data Integrity and the direct proportional of Data Confidentiality and Data Availability. According to the range of each parameter, the value of VDCI is between 1 <VDCI < 10. The minimum value of VDCI means higher level of security and the maximum value of VDCI means the lowest level of security.

After calculating VDCI value classify the data using following algorithm [3].

Algorithm
For i=1 to n
If (1< VDCI <= 3)
then R[i] =3 // Limited access Cloud is allotted to the file.
Else If (4 < = VDCI [i] <=7)
then R[i]=2 // Private Cloud is allotted to the file.
Else R[i] =1 // Public Cloud is allotted to the file.

2.2.2 Data Classification Based on Confidentiality

Munwar Ali Zardari et al. proposed the data classification technique that data is classified on the basis of confidentiality. The confidentiality is decided by the data owner or system administrator. In this data classification technique data is classified into three classes' i.e. most-sensitive, sensitive and non-sensitive. The local and public information does not require any type security that directly stored in the cloud, that types of data are considered as non-sensitive. The banking transactions, medical health record, personal data and organization data are considered as most-sensitive data. The data classification done by the data owner or system administrator, the most-sensitive data is encrypted using any encryption technique and stored in different clouds of different clusters. The sensitive data are also encrypted using any encryption algorithm, and stored into a cluster. The non-sensitive data is directly stored into a cluster, because it does not require any types of security [4].

2.2.3 K- NN Classifier

Munwar Ali Zardari et al. are proposed data classification technique based on the K-NN classifier. In this technique data is classified into two classes' sensitive (confidential) data and non-sensitive.

Sensitive (Confidential) data:
Confidential data contains very important data of individuals or organizations. The unauthorized person cannot access confidential data in the cloud. Such information is included

- Personal data: includes personal identification information such as social security number, passport number, credit card number, driver's license number.
- Financial Records: Banking transaction information, financial account number.Business Information: includes design of new product, future plan information.
- Medical/Health Data: Includes Healthcare information of person.
- Government Data: Includes government future plan information, government intellectual documents, and government agency documents.

Non-Sensitive (Non-Confidential/public) data:
These types of data are used by the general public via internet. Data which is classified as non-sensitive data includes information that is not critical to the individual or organization. Such information includes marketing material, press announcements or introductory information of an organization.

K-nearest neighbors are simple algorithm that stores all available cases that classify new cases based on a similarity measure (distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its k nearest neighbors measured by distance [5].

K-NN Algorithm:
Step 1: Determine the set of n labeled 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 neighbors based on the K-th minimum distance
Step 5: Find the classes of those neighbors.
Step 6: Determine the class of1 the new input based on a majority vote.

2.2.4 Data Classification Based on Cryptographic Parameters

Sandip K. Sood proposed a data classification technique based on the cryptographic parameters. The data stored in different sections of cloud (Public, Private and Limited access) on the basis of three cryptographic parameters (Availability, Integrity and Confidentiality). The values of three cryptographic will be gives the data owner himself and Sensitivity Rating SR will be calculated using the following formula. The value of Confidentiality (C) is based on the privacy level of data, value of Integrity (I) is based on accuracy of data and reliability of information and the value of Availability (A) is based on the how many times access data and should available in very short time when requested. After calculating Sensitivity rating SR using proposed algorithm is used to allocate the different sections in the cloud (Private, Public and Limited Access) [1].

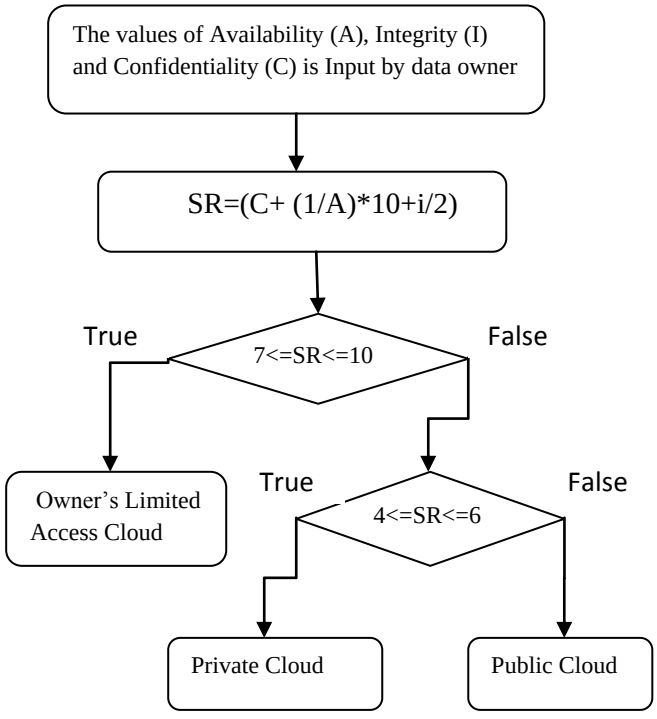

Figure3: Proposed algorithm of data classification technique

3. DATA ENCRYPTIONS TECHNIQUES

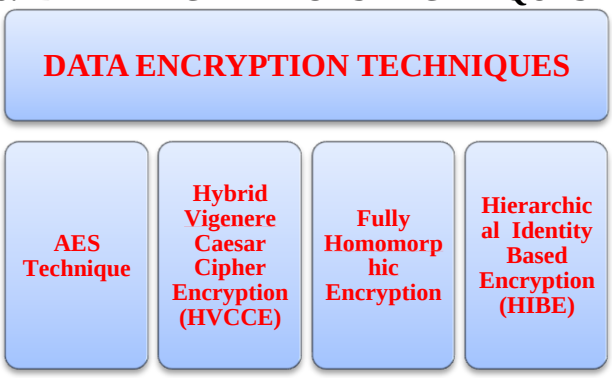

Figure4: Different Data Encryption Techniques

3.1 AES Technique

Prabhakar and Joseph have proposed a data encryption technique based on the AES algorithm. In the cloud environment AES approach protects the data from beginning to end for the entire life cycle. This encryption technique uses AES-256 algorithm for encryption and SSL (Secure Socket Layer) for protecting the data files during transfer to the cloud. The data owner or system administrator encrypts data using AES algorithm and then using SSL for upload data file security to the cloud. The proposed technique claims that provide complete security for data during all stages and it's divided into two phases. First phase deals with data encryption and upload data securely in the cloud and second phase deals with data retrieval which includes authentication of users and data decryption. In first phase data encryption is done by AES -256 encryption schemes and SSL used to securely upload data in the cloud. In second phase user need to be authenticated, the user sends the username and password to the cloud. When cloud receives the request from the user then verifies the user's details, if user is valid then start the process of data retrieval. When the user receives the data from the cloud then decrypts using AES 256 scheme. The proposed technique claims that it protects data from tampering and brute force attack. Major drawback is efficiency and privacy protection [6].

3.2 Hybrid Vigenere Caesar Cipher Encryption (HVCCE)

Nandita Sengupta and Jeffrey Holmes have proposed a new encryption technique based on the hybrid cryptography system. Hybrid Vigenere Caesar Cipher Encryption (HVCCE) is proposed for which will prevent the three cloud infrastructure like client side, server side and network. This proposed cryptographic system is designed that way the computation time for decryption of the cipher text is more compare than to any single cryptographic system for hacker. This encryption is applied on the encrypted text achieved from the second phase. Similarly decryption is to be done in three phases. In the first phase of decryption, reverse Vigenere Cipher needs to be applied with the reverse keyword on the encrypted cipher text. In the second phase of decryption, reverse Vigenere Cipher needs to be applied with the forward keyword applied on the decrypted cipher text achieved from the first phase of decryption. In the third phase of decryption, reverse Caesar cipher needs to be applied on the cipher text achieved from the second phase of decryption. In the second phase, the reverse vigenere cipher along with forward keyword is applied on encrypted data files achieved from the first phase of decryption. In the third phase, the reverse Caesar cipher is applied to the encrypted data files achieved from the second phase of decryption and get the plain text to the user. The proposed technique main advantage is providing triple encryption to the data and major drawbacks is less efficient [7].

3.3 Fully Homomorphic Encryption

Feng Zhao, Chao Li and Chun Feng Liu have proposed a new kind of data security is based on Craig Gentry construct a homomorphism encryption scheme including 4 methods. The four methods are key generation algorithm, encryption algorithm, decryption algorithm and additional evolution. In the homomorphic encryption scheme, it can apply a mathematical manipulation to the plain text for encryption. In homomorphic encryption technique support only one operation either multiplication or addition but fully homomorphic encryption technique supports both operations like addition and multiplication. The proposed technique, it can apply any number of addition algorithm and the multiplication algorithm to encrypt the plain text. The proposed security solution is fully rigorous for retrieval and processing of encrypted data. The fully homomorphic encryption technique has proposed new encryption algorithm. In encryption algorithm, the encryption parameters p, q and r, p are positive odd integer, q is a large positive number, and r is random number selected when encryption. The p and q determined in key generation phase, the encryption done by following formula and m is a plain text message.

$$c= m+ 2*r + p*q$$

We get the cipher text using the above formula, and then upload the cipher text to the cloud. When data retrieval from the cloud, in decryption algorithm manipulating some mathematical operation on cipher text to get the plain text. The proposed decryption algorithm is following

$$m=(c \bmod p) \bmod 2.$$

We get the plain text messages by using above formula. The advantage of this proposed technique is effectively leading to the broad prospect, the security of data transmission and the storage of cloud computing. The main disadvantage is at present, fully homomorphic encryption scheme has a high computation problem needs further study [8].

3.4 Hierarchical Identity Based Encryption (Hibe)

Xin Dong et al. have proposed encryption technique "SECO" which is secure and efficient collaboration scheme based on Hierarchical Identity Based Encryption (HIBE). The proposed encryption technique is to ensure the data confidentiality on the untrusted clouds.

Hierarchical Identity Based Encryption (HIBE) is an encryption technique that is used to control users who are unauthorized or partially authorized users and might be share private key to unauthorized user which wills the lead unauthorized data access. Hierarchical Identity Based Encryption has five steps: setup, encrypt, key gen, decrypt and delegate. In setup step it takes security parameters as input and gives output is the master key, in encrypt step it takes plain text, identity vector and public parameters as input and output as cipher text. In key gen step it takes the master key, identity vector and public parameters as input and output as secret key for public vector, in decrypt step takes cipher text, secret key and public parameters as input and output as plain text. In delegate step takes the secret key for identity vector,

identity and public parameters as input and output as secret key for identity, it concatenation of identity vector and identity.

The proposed technique uses two levels of Hierarchical Identity Based Encryption to ensure confidentiality of data files in untrusted clouds. The proposed encryption technique first explores the secure data collaboration service that prevents information leakage and enables one –to- many encryptions. It also enables the fine –grained access control and data writing simultaneously. The proposed technique provides the data collaboration service that supports the consistency and availability of the shared data among the multi-users. The proposed encryption technique employs a two level HIBE scheme, which contains Private Key Generator is trusted third party which assigns the secret keys to the Domain- Private Key Generator. Root Private Key Generator manages the independent co-operative Domain Private Key Generators while Domain Private Key generator manages the end users. Root Private Key Generator generates the master key and private keys for the Domains. The Data owner wants to encrypt the data, it uses Public key of multiple recipients so that only deliberated domains can decrypt that data. The user is will decrypt the data, it request for secret key to the domain Private key generator and Domain Private Key Generator providing the secret key for decryption of data. The main advantage of proposed technique SECO is highly efficient and low overhead on computation and communication [9].

4. COMPARISON STUDY

Table 1: Comparison Studies Of existing data encryption Techniques

Researcher	Encryption Techniques	Advantages	Limitations
Prabhakar and Joseph	AES	The proposed technique claims that it protects data from tampering and brute force attack.	Major drawback is efficiency and privacy protection
Nandita Sengupta and Jeffrey Holmes	Caesar cipher and Vigenere Cipher	The proposed technique main advantage is providing triple encryption to the data.	A major drawback is less efficient.
Feng Zhao Chao Li and Chun Feng Liu	Craig Gentry construct homomorphism encryption scheme	The advantage of this technique is effectively leading to the broad prospect, the security of data transmission and the storage of cloud computing.	The disadvantage is at present, fully homomorphic encryption scheme has a high computation problem needs further study
Xin Dongy, Jiadi Yuy, Yuan Luoy, Yingying Chenz, Guangtao Xuey and Minglu Liy	Hierarchical Identity Based Encryption (HIBE)	The main advantage of proposed technique SECO is highly efficient and low overhead on computation and communication.	Collusion Resistance and user accountability.

5. CONCLUSION

The cloud computing is one of emerging paradigm, and security of data in the cloud is the most important issue which acts barrier in the implementation of cloud computing. In this paper survey on the existing data classification techniques used in cloud computing. The objective of data classification is to establish the required level of security for data and to protect data by providing a sufficient level of security according to the risk levels of data. In this paper also survey on the existing encryption techniques, that protect the data for the entire life cycle from the beginning to the end in the cloud computing. The brief survey also provides a comparison studies of existing encryption techniques used in cloud computing. In comparison study discuss their advantages and limitations of different data encryption techniques. The Fully Homomorphic Encryption and Hierarchical Identity Based Encryption (HIBE) are best security techniques for data on cloud environment, because it tackles maximum security issues in cloud computing. In future work, aim that to propose a scheme that will contain the security features in these while overcoming disadvantage and open issues in them.

6. REFERENCES

[1] S. K. Sood, 2012 A combined approaches to ensure data security in cloud computing, ACM, Journal of Network and Computer Applications, vol. 35, no. 6, pp. 1831–1838.

[2] Peter Mell and Timothy Grance, 2011.The NIST Definition of Cloud Computing,NIST Special Publication 800-145

[3] Faraz Fatemi Moghaddam, Moslem Yezdanpanah ,Touraj Khodadadi 2014 VDCI: Variable Data Classification Index to Ensure Data Protection in Cloud

Computing Environments, IEEE Conference on Systems, Process and Control (ICSPC 2014), pp.53-57 2

[4] Munwar Ali Zardari, Low Tang Jung, Nordin Zakaria, 2013 Hybrid Multi-cloud Data Security (HMCDS) Model and Data Classification IEEE Advanced Computer Science Applications and Technologies (ACSAT), pp. 166-171 2

[5] Munwar Ali Zardari, Low Tang Jung, Nordin Zakaria, 2014 K-NN Classifier for Data Confidentiality in Cloud Computing, IEEE Computer and Information Sciences (ICCOINS), pp. 1 – 6

[6] D.M. Prabhakar, K.S. Joseph, 2013, A new approach for providing data security and secure data transfer in cloud computing**,** International Journal of Computer Trends and Technology (IJCTT) pp 1202-120

[7] Nandita Sengupta, Jeffrey Holmes 2013, Designing of Cryptography Based Security System for Cloud Computing, IEEE International Conference on Cloud & Ubiquitous Computing & Emerging Technologies pp. 52-57

[8] Feng Zhao, Chao Li, Chun Feng Liu 2014, A cloud computing security solution based on fully homomorphic encryption, IEEE Advanced Communication Technology (ICACT), pp. 485 - 488

[9] Xin Dongy, Jiadi Yuy, Yuan Luoy, Yingying , Guangtao Xuey and Minglu Li, 2013, Achieving Secure and Efficient data collaboration in Cloud computing, IEEE, pp1-6