A comprehensive Study of Data Masking Techniques on cloud

K.Sharmila
Assistant Professor & Research Scholar,
Dept. of Computer Science,
Vels University, Chennai,
Tamilnadu,
aswinpalani14@gmail.com

S. Borgia Anne Catherine
Assistant Professor
Dept. of Computer Science,
A.M Jain college, Chennai,
Tamilnadu.
sreeja.vs87@gmail.com

Sreeja V.S
Assistant Professor & Research Scholar,
Dept. of Computer Science,
Vels University, Chennai,
Tamilnadu, India

ABSTRACT
This paper discusses about Cloud computing which is a technology that provides low cost, scalable multiplicationsability and services to enterprises on demand for growth. Although, cloud computing is facilitating the Information technology industry, the research and development in this field is yet to be satisfactory. Cloud computing resources offered service on an as-needed basis and delivered by IP-based connectivity, providing highly scalable, reliable on-demand services with agile management capabilities. Security in data on cloud has become one of major aspects in this era. Cloud security is a set of security principles which can be applied to protect the data from the unauthorized users which are not reliable. Thereby ensuring reliability and trust in the cloud environment can be employed using different data masking Techniques. Data masking is the process of protecting sensitive data from thefts and hackers. This paper gives a detail study of different models on cloud and data masking techniques which can be integrate to secure the data.

Key words: Cloud Computing, cloud Security, Data Masking.

2. INTRODUCTION
Cloud Computing is a technology which uses the internet and central remote servers to maintain data and applications. It is a group of computers and servers linked together over the internet. It refers to manipulate, design and access the applications online. It allows consumers and business men to use applications without installation and access their personal file from any computer with the help of internet. [1] It also offers online data storage, infrastructure and application. It is architecture for providing computing services via internet on demand and pay per user access to a pool of shared resources for the network storage, services and applications. It is totally an internet based technology in which client data is stored and maintained in data center of cloud provider like Google, Amazon, Salesforce.com and etc. The resources in cloud system are transparent for the application and the client do not know the place of resource. The client can access your application from anywhere. The amount of resources provided in the cloud system for the client is increased when their requirements are high and decreases when their requirements are less. The cloud computing can be seen as the important change of information industry and will make more impact on the development of information technology for the society.

The following are some of the notable challenges associated with cloud computing and although some of these may cause a slowdown when delivering more services in the cloud, most often can provide opportunities, if resolved with due care and attention in the planning stages are

- Security and Privacy
- Lack of Standards
- Continuously Evolving
2. CLOUD SERVICES MODELS

There are three types of cloud computing services models:

**SaaS (Software as a service)**

Software as a service (SaaS) is a cloud computing offering that provides users with access to a vendor’s cloud-based software. Users do not install applications on their local devices. Instead, the applications reside on a remote cloud network accessed through the web or an API. Through the application, users can store and analyze data and collaborate on projects (3).

**PaaS (Platform as a service)**

Platform as a service (PaaS) is a cloud computing offering that provides users a cloud environment in which they can develop, manage and deliver applications. In addition to storage and other computing resources, the users are able to use a suite of prebuilt tools to develop, customize and test their own applications. PaaS provides a platform with tools to test, develop, and host applications in the same environment.

**IaaS (Infrastructure as a service)**

Infrastructure as a service (IaaS) is a cloud computing offering in which a vendor provides users access to computing resources such as servers, storage, and networking. Organizations use their own platforms and applications within a service provider’s infrastructure.

![Cloud Computing Service Models](image_url)

Fig 1. Cloud Computing Service Models
3. CLOUD COMPUTING DEPLOYMENT MODELS

Cloud hosting deployment models represent the exact category of cloud environment and are mainly distinguished by the proprietorship, size and access. It tells about the purpose and the nature of the cloud. Most of the organizations are willing to implement cloud as it reduces the capital expenditure and controls operating cost. In order to know which deployment model matches your website requirements it is necessary to know the four deployment models.

1) Public Cloud: It is a type of cloud computing model in which service provider provides resources to the general public over the internet.

2) Private Cloud: It is a type of the cloud computing models which provides the same features same as that of the public cloud but it is dedicated to only single organization.

3) Hybrid Cloud: It is combination of both private and public cloud service which performs distinct functions within same organization.

4) Community Cloud: It is a type of cloud hosting in which the setup is mutually shared between many organizations that belong to a particular community, i.e. banks and trading firms.

4. DATA MASKING

Data Masking is a method to hide sensitive information. As people interchangeably use such terms as data anonymization, data de-identification, data scrambling, data scrubbing, and data obfuscation, understanding which term defines which process gets confusing. Moreover, oftentimes, there are subtle differences in how this or that vendor defines the process. Data masking is not just a science of algorithms. It is also a science of public data sets.
Knowledge about public data sets definitely requires training for the person implementing the solution, together with understanding of k-anonymity and l-diversity concepts.

Fig/3. Data Masking

Types Of Data masking

There are different ways to mask data. Different methods are used for different needs, with different organizational roles, and in different scenarios.

**STATIC DATA MASKING (SDM)**

Using stable non-changing environment that originally has a copy of production database, often times called "golden copy" to anonymize data set is called "static data masking". Mainly used to refresh non-production environments and prevent insider’s threat.

**DYNAMIC DATA MASKING (DDM)**

Certain categories of employees, based on their role in the organization, should have access only to a part of the record or to a whitened values in the application. Such categories examples might be customer service
representatives who do not have access to the credit card besides the last four digits, or judicial clerks who should not have access to addresses and names.

Examples of a masked credit card number data is:

4XXX-XXXX-XXXX-5558

Different types of data masking techniques

**SHUFFLE**

Shuffle is a data masking algorithm that allows to preserve all the values in the given column as they were while it changes the position of the value(s) in the column comparing with their initial position. The name reflects a "shuffling" action. The algorithm is useful when it is necessary to preserve the aggregated values and it could be used for columns with unique constraint. As an example, let's consider a column with sales numbers - if the resulting quarterly sales need to remain intact so that not to break an application, we would want to use a shuffling algorithm.(5)

**SUBSTITUTION**

The substitution technique replaces the existing data with random values from a pre-prepared data set. i.e. this technique consists of randomly replacing the contents of a column data with information that looks similar but is completely unrelated to the real details. For example, the surnames in a customer\ database could be sanitized by replacing the real last names with \surnames drawn from a largish random list. Substitution is very effective in terms of preserving the look and feel of the existing data.(thesis)

**RANDOM**

Random substitution masks data by replacing a given value with a random value from a pre-compiled data set. Values in the data set are suitable and conform to the same rules or definitions that of the given value. With each iteration of the masking algorithm, another random value gets chosen and replaces the value on the input. An example would be a value of John being masked with values of Alex, Robert, and Mike in subsequent iterations of the substitution operation.

As data masking of the given value with Random Substitution does not guarantee to be repeatable among cycles, it is not suitable for masking data sets with unique constraints and requires extra work of preserving referential integrity if used on the values of fields preserving referential integrity constraint. In denormalized data sets it might cause difficulties in implementing row-internal synchronization, table internal synchronization and table-to-table synchronization operations.(5)
NUMBER and Date VARIANCE

The number and Date Variance technique varies the existing values in a specified range in order to obfuscate them. For example, birth date values could be changed within a range of +/-60 days. The Number and Variance technique is useful on numeric or date data. The algorithm involves modifying each number or date value in a column by some random percentage of its real value. (6)

NULLING Out

The Nulling out technique simply removes the sensitive data by deleting a column of data and replacing it with NULL values is an effective way of ensuring that it is not inappropriately visible in test environments. Unfortunately it is also one of the least desirable options from a test database standpoint. (6)

Encryption:

Encryption is one of the most complex methods to solve the data masking problem. The Encryption technique algorithmically mix-up the data. This usually does not leave the data looking realistic and can sometimes make the data larger. Encryption also destroys the formatting and look and feel of the data. Encrypted data rarely looks meaningful; in fact, it usually looks like binary data. This sometimes leads to character set issues when manipulating encrypted varchar fields. Certain types of encryption impose constraints on the data format as well. [4]

Based on detail study ,the usage of different method used across the industries, we have to find which would be far superior in terms of performance and data security. We have to apply different size of data sets for verification. On the whole in various industries based on the performance we would find which Data Masking technique is to be used for data security.

5. CONCLUSION

In this paper we discussed about the cloud services models, deployment models and security in cloud by using data masking techniques. Storage of data on the cloud refines the way we manage the storage of data and access the data from the cloud. This paper also explores the need of data masking in present information. Data masking will enable us to accomplish the following: (a) Increase protection against data theft. (b) Enforces ‘need to access’. (c) Provides realistic data for testing, development and data sharing. (d) Provides a heightened sense of security to clients, employee and supplier. In future the extend of analyzing the data masking techniques further with the various analytical parameters across industries has to be evaluated.

REFERENCE


11. Design of Data Masking Architecture and Analysis of Data Masking Techniques for Testing Ravikumar G K, Dr. B. Justus Rabi, Dr MGR University, Chennai, Tamil Nadu, INDIA