Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 493 – 498

Procedia Computer Science

Data Classification for achieving Security in cloud computing

Rizwana Shaikh[a], Dr. M. Sasikumar[b]

[a]*Assistant Professor, SIES Graduate School of Technology, Nerul. Navi Mumabi*
[b]*Associate Director, CDAC Kharghar, Navi Mumabi, India.*

Abstract

Data is the valuable asset and of great concerns when moving towards the cloud. Data privacy and security is the active area of research and experimentations in cloud computing. Data leakage and privacy protection is becoming crucial for many organizations moving on to cloud. Data can be of various types and degree of protection required for all the data is also varies. Here we are proposing a classification technique that defines various parameters. Parameters are defined based on various dimensions. Data security can be provided based on the level and the required protection. Corresponding security provisions at the storage can be applied based on data set classified as per the dimensions. The efficiency of the proposed classification scheme is analyzed with the sample dataset collected.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

Keywords: Cloud Computing; Data; Data Classification; Security;

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.087

1. Introduction

Data is a vital asset for any organization. Data could be in any forms, i.e. numbers, words, images etc. Data privacy and security is a crucial issue for any organization. Data deals with various properties such as accuracy, validity, reliability etc. described below.

- Accuracy: It deals with data correctness defined by the original source. Data should be accurate for the intended use and should be captured only once, although it may have multiple uses. Data is captured at the time of activity.
- Validity: Data that is recorded and used with respect to relevant requirements. It should be valid over a period of time.
- Relevance: Data captured should be used with respect to relevant requirements.
- Completeness: Data should be complete with respect to its usage.
- Accessibility: It deals with the access of data with respect to time and cost.
- Consistency: It deals with the uniformity of the content with respect to changes and transactions that uses data.

Basic security issues of data include confidentiality, integrity and availability. Data confidentiality deals with privacy of data which includes authentic and authorized access of sensitive data. Data integrity deals with the data content. Consistency and accuracy of data is required to achieve integrity. Data availability issues pertaining to foolproof storage, storage type, provisions for disaster recovery and backup plan.

Data availability concern is a vital and critical issue to any organization moving to the cloud. Security issues with respect to data increases when moving to cloud. User controls about data, data protection mechanism provided and data availability are some of the issues that user needs to know before using cloud for data storage. The data stored on the cloud should be protected from unwanted malicious disaster. This disaster could be man-made or natural. A cloud provider should be aware and accordingly provide measures to achieve data availability at all the times.

Data Classification is the process of defining various data levels and deciding a level of sensitivity to it. It is an essential activity at various stages as it is being created, modified, stored, or transmitted. The classifications of the data determine the extent to which the data needs to be secured and its value in terms of Business Assets. Data classification is done based on the various aspects. Some classify the data according to the risk associated with the disclosure. They are public, internal, confidential (or highly confidential), restricted, regulatory, or top secret. Some classify the data based on the way it is created, user personal data, their usage patterns etc.

In a cloud computing environment data asset is very crucial depending on the business and the service delivery models. To provide the controlled access and authorization, classifying data based on security level criteria becoming area of interest by many organizations using or providing cloud services. Here we have studied a set of classifications given in the literature and identified a set of parameters based on the security requirements for cloud data. We have analyzed some data sets that can be used to provide the security based on their usage and access control with respect to cloud computing environment.

2. Literature Review

Data privacy and security concerns in cloud computing is always an issue. Storage and access mechanism proposed by various researchers and experimentations shows that inspite of having provisions for data security, various attacks and data leakage problems and still part of the cloud ecosystem. Here we have identified some of the research exist in the literature. Outsourcing sensitive data to a cloud service provider with the permission of block level modifications is proposed in [1]. Indirect mutual trust is established between the data user and owner by using third party. Pearson in [2] discusses policies and assessment procedures for privacy enhancement methods and tools. Privacy in terms of legal compliance and user trust, data leakage for sensitive data are provided. Authors in [3] gave a benchmark to secure data-in-transit in the cloud. Protecting data during migration is discussed via benchmark for encryption overhead and security. More encryption is desirable for strong security but it requires more computation. So a benchmark gives balance for the security and encryption overhead. Large scale search system for the purpose of Information exchange between internet communities leads to formation of covert Channels [4]. An agent based security model to control data from covert channel is presented. It may solve the problem of data leakage in the cloud environment. Authors in [5] discuss the privacy issue by retaining data control to user to increase confidence. Cloud computing attacks are discussed and some provisions and means to overcome from the same are proposed. A novel patient-centric framework and a suite of mechanisms for data access control of Patient Health Record, is presented by

the authors in [6]. Data is stored in semi-trusted servers. Attribute based encryption techniques to encrypt each patient's Health Record file, is used to achieve fine-grained and scalable data access control.

Data characteristics are analyzed with respect to online social networks by authors in [7]. They have identified the information and their leakage at the time of sharing data in the network. They are also using third party check for this purpose. Data standards and transmission controls are studied by the author to provide privacy in the social networks. Data protection taxonomy by considering various aspects is proposed by the author in [8]. A case study is examined by providing solution for data protection in terms of questions like who requires protection, what is to be protected etc. A solution for the industry related information is proposed and discussed. A restriction for providing data accessed by the third party applications is proposed in the form of framework in [9]. Policies are applied for restricted access and thus achieving the privacy of user confidential data from the installed third party applications. A privacy concerns has brought to the notice of the user that makes use of hotspots for the internet access in [10]. Real data has been analyzed and proposed the security risk associated with it. Various categories of data are observed based on the types of the privacy concerns. A three dimensional view for data taxonomy is proposed in [11]. It classifies data as per visibility, Granularity and purpose. Various levels are defined along these dimensions to achieve the data privacy. Taxonomy for social data is presented by the author in [12]. It classifies the data based on the way it is generated in the social network, and accordingly the privacy and access rights have to be applied. Data classification at various stages in a social network is given in [13]. It classifies data based on security parameter confidentiality i.e. the data disclosure in a network. It applies this classification in various phases of data like collection, processing, dissemination and invasion.

Data security is studied as a part of survey by authors in [14]. Various other security issues are also analyzed and a trust based solution for the same is proposed. Here we have identified data security concerns and attempting to provide security by classifying the data. Various dimensions are identified along which security and protection level can be applied to different data with varying degree of values.

3. Proposed Data classification

Data classification is the process of identifying data elements with respects to its value in the business. Value is identified based on their usage and access control restrictions. Figure1 indicates the three types of characteristics on which data has to be classified and accordingly security considerations can be applied.

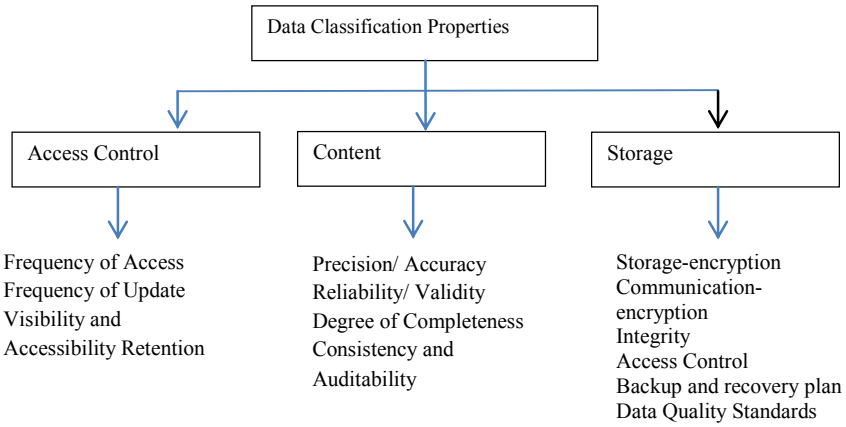

Fig.1 Data Classification in cloud computing

3.1. Access Control

This category defines the access restrictions applied on data. It includes;

- Frequency of access: Data elements can be accessed more frequently, less frequently or moderate number of times. A user can decide the threshold or maximum limit for these ranges and can classify them giving one of the three values.
- Frequency of update: Updation of data can also be performed repeatedly. It yields the vale less, moderate or more as above.
- Visibility and Accessibility: The data can be classified based on the accessibility and visibility region. It can take the value restricted with respect to some criteria or to all. Criteria for restriction can be determined by the data owner and the organization usage of it.
- Retention: One of the parameter for classifying data could be the retention period for data availability in the system.

3.2. Content

Content of data possess properties with respect to its modifications. Data content possess several properties and can be classified as below.

- Precision/ Accuracy: Accuracy of the data can be used to classify it as high, low or poor. The content of high precision and accuracy is desirable for some data elements over the other.
- Reliability/ Validity: Depending on the accuracy, reliability and validity of the data can be determined. It can take the value as low, medium and high.
- Degree of Completeness: For some data elements, degree of completeness can be used to classify. It could be mandatory or optional otherwise for the selected data for completeness.
- Consistency: Data consistency property describes data accuracy at any point of time. For some data consistency is must, while for some cases it is not required at all. For that data once stored is becomes permanent storage. No updation can be possible at all for such data elements.
- Auditability: As with respect to consistency, some data are auditable and other is not. This makes auditability possible or no to classify the data items.

3.3. Storage

Data storage policies can be applied based on the criteria and constraints applied to the different types.

- Storage-encryption: Encryption of the data based on the size of encryption key. As the security strength required for the data increases, it will require large size key. As the key size is more time require to break the key is more hence more security. Hence a benchmark is selected as per the security and computational overhead with respect to data.
- Communication-encryption: Data moving to or from the system also prone for leakage and eavesdropping. A communication encryption should be provided for sensitive and restricted data items.
- Integrity: The data integrity is critical issue and has to be addressed by hash algorithm available like MD5, SHA, etc. It also applied based on the security level required to be achieved for the specific data elements.
- Access Control: A predefined access control policy has to be associated with the various data elements. Role based access control for various user and privileges has to be defined based on the policies and restrictions guidelines.
- Backup and recovery plan: Backup plan for the storage is essential requirement for disaster and recovery purpose. So based on the criticality of the data a backup plan should be associated.
- Data Quality Standards: Various standards for certifying data are also desirable by the user at the time of classification of data. A data quality standard increases the security of the stored data in the system.

  The above classification scheme can be used to provide various degrees of security for data. Data elements can be tagged at the time of storage. Based on the tag required security can be provided to that data element.

4. Analyzing data classification

Data classification approach can be studied and the effectiveness of it can be determined by simulating on a sample data set. Simulating the classification of data based on personal data set is analyzed. Personal data elements like name,

addresses etc. are taken as the sample. We have used the subjective criteria to classify them and accordingly security provisions for the storage and communications can be incorporated. These criteria can also be converted to values and threshold can be set for objective evaluation. Data classification is shown in table1 where all parameters are used. Individual parameters are classified by using subjective criteria.

Table1: Data classification: Man= Mandatory; R= Restricted; Mod= Moderate; FU= Frequency of Update; FA= Frequency of Access; CE= Communication Encryption; SE= Storage Encryption; Int.= Integrity; Con/Aud=Cosistency/ Audibility; DoC= Degree of Completeness; R/V= Reliability/Validity; P/A= Precision /Accuracy

	Properties related with General security				Properties related with data content				Properties require for cloud Storage			
Data Elements	FA	FU	Visibility	Accessibility	P/A	R/ V	DoC	Con./ Aud	SE	CE	Int.	AC
Name	More	Never	All	All	High	High	Man	Yes	Less	Moderate	Less	No Control
Address	Moderate	Less	R	R	High	High	Man	Yes	Strong	Strong	Strong	RBAC
Phone	Moderate	Less	R	R	High	High	Man	Yes	Strong	Strong	Strong	RBAC
Mobile	More	less	R	R	High	High	Optional	Yes	Strong	Strong	Strong	RBAC
DOB	Less	Never	R	R	High	High	Man	NA	Strong	Strong	Strong	RBAC
Place of Birth	Less	Never	R	R	High	High	Optional	Yes	No	Less	No	RBAC
Caste	Less	Never	R	R	High	High	Optional	Yes	Strong	Strong	Strong	RBAC
Sex	Moderate	Never	All	All	High	High	Man	NA	Less	Moderate	Less	No Control
Nationality	Less	Never	All	All	High	High	Man	NA	No	Less	No	No Control
Blood Group	Less	Never	R	R	High	High	Optional	Yes	Less	Moderate	Less	RBAC
Height	Less	Never	R	R	High	High	Optional	Yes	Less	Moderate	Less	RBAC
Marital status	Moderate	Less	All	All	High	High	Man	Yes	Less	Moderate	Less	No Control
Passport no	Less	Less	R	R	High	High	Optional	Yes	Less	Moderate	Less	RBAC
SSN	Moderate	Never	R	R	High	High	Man	NA	Strong	Strong	Strong	RBAC
Email	More	Less	R	R	High	High	Man	Yes	Strong	Strong	Strong	RBAC
Physical Characteristics	Less	Never	No-one	No-one	High	High	Optional	Yes	Moderate	Strong	Moderate	No-one
Eye Color	Less	Less	No-one	No-one	High	High	Optional	NA	Strong	Strong	Strong	No-one
Biometric	Less	Less	No-one	No-one	High	High	Man	NA	Strong	Strong	Strong	No-one

Backup and recovery plan for the data and the standard depends on the application and the organization that possess or requires the standard for quality storage. Data can be classified and based on the type of accessibility and content storage provisions can be adapted to provide the security.

5. Conclusion

Data privacy and security is one of the major issues while dealing with the data storage in cloud. Many classification techniques exist in the literature that classifies the data in social network or other application area. We have identified a set of parameters for data classification in cloud. It is for providing security levels based on type of content and accessibility. We are providing the level of security in cloud storage as per the required confidentiality and access restrictions for the data specified. We have analyzed few data elements and classified them based on the

proposed parameters. All the elements that are stored in cloud storage can be classified first based on the content and access control parameters. Based on that, classification provisions can be given for storage and communication encryption, integrity and access control mechanisms. Also a regularized backup plan can be decided for disaster and recovery. Data security or quality standard improves the strength substantially. The proposed classification can be implemented as a working module i.e. prototype and simulation of the classification technique can be evaluated.

References

1. Ayad Barsoum and Anwar Hasan, "Enabling Dynamic Data and Indirect Mutual Trust for Cloud Computing Storage Systems", IEEE Transactions on Parellel and Distributed Systems, Dec. 2013 (vol. 24 no. 12), pp. 2375-2385.
2. Pearson S, "Taking account of privacy when designing cloud computing services", Software Engineering Challenges of Cloud Computing, pages, 44 – 52, Vancouver, BC, 2009.
3. Ji Hu and Klein A, "A Benchmark of transparent data encryption for migration of web application in cloud", Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pages 735-740, Chengdu, 2009.
4. Tetsuya M, Kazuhiro S and Hirotsugu, K., "A system for search, access restrictions and agents in the Clouds", Ninth Annual International Symposium on Applications and the Internet Cloud, Pages 201-204, Japan, 2009.
5. Descher M, Masser P, Feilhauer T, A Min Tjoa and Huemer D, "Retaining data control to the Client in Infrastructure Cloud", International Conference on Availability, Reliability and Security, pages 9-16, Dornbirn, 2009.
6. Ming Li, Shucheng Yu, Yao Zheng, Kui Ren and Wenjing Lou, "Scalable and Secure Sharing of Personal Health Records in Cloud Computing using Attribute-based Encryption", IEEE transaction on parallel and distributed systems, pages 131-43 vol. 24, issue 1, 2012.
7. Balachander Krishnamurthy and Craig E. Wills, "Characterizing Privacy in Online Social Networks", Proceedings of the first workshop on Online social networks, WOSN '08, Pages 37-42, ACM New York, 2008.
8. Mike Dutch, A Data Protection Taxonomy, Storage Networking Industry Association, June 2010.
9. Yuan Cheng, Jaehong Park and Ravi Sandhu, Preserving User Privacy from Third-party Applications in Online Social Networks, Proceedings of the 22nd international conference on World Wide Web Companion, Pages 723-728. Geneva, Switzerland, 2013.
10. Ningning Cheng, Xinlei (Oscar) Wang, Wei Cheng, Prasant Mohapatra, Aruna Seneviratne, Characterizing Privacy Leakage of Public WiFi Networks for Users on Travel, IEEE International Conference on Computer Communications, Italy, 2013.
11. Ken Barker, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam Majedi, Sampson Pun, and Adepele Williams, A Data Privacy Taxonomy, Advanced Database Systems and Applications Laboratory, University of Calgary, Canada, 2009.
12. Bruce Schneier, A Taxonomy of Social Networking Data, The IEEE Computer And Reliability Societies, August 2010.
13. Sergio Donizetti Zorzo, Rodrigo Pereira Botelho, Paulo Muniz de Ávila, Taxonomy for Privacy Policies of Social Networks Sites, Published Online, Social Networking, 2013, 2, 157-164 October 2013 (http://www.scirp.org/journal/sn).
14. Rizwana Shaikh and Dr. M. Sasikumar, "Security Issues in Cloud Computing: A survey. International Journal of Computer Applications 44(19):4-10, April 2012.