Secure Data Access

1. Introduction

The popularity of online shopping has led to a large amount of consumer data being collected, stored and managed by vendors and/or third-party service providers. How to protect the identity privacy of consumers while achieving secure access control has become an urgent problem. This report will consist of four parts. The first part presents the assumptions on which the experiments are based; the second part gives definitions and explanations of some terms; the third part will present an implementation of the system workflow; the fourth part will discuss whether it fulfils the task requirements and explore possible limitations and improvement method.

2. Assumptions Clarification

This report is based on the following assumptions：

1. Each vendor manages only its own consumer data.

2. All vendors have a centralized trusted third party (C-TTP). C-TTP issues access credentials to different user groups.

3. You have access to symmetric (e.g. AES) and public-key (e.g. RSA) cryptosystems and hash functions (e.g. MD-5).

4. Users (i.e., data requesters) can be assigned to one of three user groups (G1, G2, and G3). Each group is granted a different level of access rights defined in L1, L2, and L3. L1 access privilege encompasses the rights of L2 and L3. Similarly, L2 access privilege encompasses the right of L3.

3. Definitions

Attribute certificate. An attribute certificate (AC) has a data structure that is almost equivalent to that of an identity certificate (IC). However, the main difference is that an attribute certificate does not contain a public key, but rather attributes that specify access control information associated with the AC holder, including group membership, roles and security clearance. (Mavridis et al., 2001) Attribute authentication provides a simple and consistent means to extend the identity-based public key infrastructure(PKI) to support role-based authorization policies. (Linn et al., 1999). The validity period of general attribute certificates is relatively short, which can avoid the problems of public-key certificates in handling CRL. If the validity period of the attribute certificate is very short, the certificate will automatically expire when the validity date is reached, thus avoiding the disadvantages of the public key certificate when revoking.

Privilege Management Infrastructure. The Privilege Management Infrastructure (PMI) uses AC to check and validate authorization data for a resource's certificate subject. "While PKI uses Public-Key Certificates (PKCs) to prove the identity of a certificate subject, PMI uses ACs to check and validate authorization data to a certificate subject for a given resource." (Gergely et al, 2016). The structure of an AC can be shown as follows(Chadwick et al, 2004) :

- the version number of this AC (v1 or v2)
- identification of the holder of this AC
- identification of the AA issuing this AC
- the identifier of the algorithm used to sign this AC
- the unique serial number of this AC
- the validity period of this AC
- the sequence of attributes being bound to the holder
- any optional extensions

Pseudonymization. Pseudonymization is a way to achieve de-identification. It means that real identifiers are replaced by pseudonyms that are unique to the individual but unrelated to the person in the "real world". There are two types of Pseudonymization, either one-way pseudonyms that cannot be reversed, or reversible pseudonyms that enable patient re-identification. (Noumeir et al., 2007) . **One-way pseudonym** can be implemented with the help of a hash function, which by our assumption should be MD5. **Reversible pseudonyms**, on the other hand, can be implemented by either saving a mapping between user identifiers or pseudonyms or by saving the parameters of reversible mapping functions that can be applied. (Noumeir et al., 2007) Therefore, the conversion between pseudonyms and real names can be achieved either by maintaining a database containing mapping relations between pseudonyms and real names or by maintaining a key and using an encryption algorithm, which by our assumption should be AES, as a mapping function.

4. Method Design and Explanations

4.1. (a)

The design and workflow of the entire system are shown in the Fig. 1 and can be described as follows:

Firstly, the Client applies for an Authentication and Authorization certificate.

1. Firstly, the client applies to PKI for an identity certificate. Specifically, the user generates his public key and private key based on the RSA algorithm, and generates a certificate signing request(CSR) by a public -key and subject identifier, submits it to Certificate authority, CA reviews and signs the certificate, and finally returns it to the user.
2. Next, the client applies to PMI for an attribute certificate. The overall process of application would be virtually identical to PKI's, except for some changes on their names, like the certificate issuer, which is CA in PKI, should be attribute Authority (AA) in PMI.

Second, the client establishes a secure channel with the reverse proxy host.

1. The first node which client interacts with is the bastion host, this bastion host will take the following roles. Firstly, to achieve confidentiality, it will be responsible for the role of a reverse proxy, interacting with the user on behalf of the whole system. This is reflected in the fact that the entire distributed data system is externally opaque, so that when the user queries the data he cannot know which vendors the data comes from. Second, to protect the internal network security, the firewall policy will be in effect here.

2. The client sends his IC, and the proxy host, after verifying it, will use the clients' public key to negotiate with the client about the symmetric encryption algorithm, which in this case will be AES, and the master secret K to be used. The final transmitted data is first hashed with MD5, generating MAC. And the MAC is attached to the end of the data which is encrypted with AES, $Msg = AES(data, K)||md5(data)$.

Third, the client sends a data request message encapsulating its AC, and the proxy host passes the message to the data host in the intranet for the query.

1. When the data request message encapsulating the AC arrives at the data host, the host extracts the role and group information in the AC and confirms the privileges of the group the user is in through the Role-Based Access Control (RBAC). The data host will perform a single query for the distributed clusters of all vendors and generate the query structure into three de-identified record groups. Specifically：

    1. Generating de-identified **record 1** by using the Reversible pseudonyms method. The data host maintains a key database, selects the associated key k when performing reversible pseudonymization, and performs a single mapping of data from all vendors by a symmetric encryption algorithm, such as AES, $p_{uid} = AES(k, uid)$. This Synchronized mapping ensures that data from different vendors can be linked by the same pseudonym ID. Also, the same k value is always used to generate a fixed p_{uid} when using the same pseudonym as the index of a query, thus the effect of continuous tracking of a specific user can be achieved.

        With RBAC, record 1 can only be accessed by users who have L1 privilege.

    2. Generating de-identified **record 2** by using the One-way pseudonym method. The data host maintains a random number generator that generates random numbers as HMAC salt when performing One-way pseudonym, and performs a single mapping of data from all vendors, $p_{uid} = MD5(salt, uid)$. This synchronized mapping approach ensures that data from different vendors can be linked by the same pseudonym ID. However, the value of the generated p_{uid} is not fixed due to the different random numbers generated, thus achieving the effect of anonymity.

        With RBAC, record 2 can be accessed by users who have L1 or L2 privilege.

    3. Generating de-identified **record 3** by using the One-way pseudonym method. The data host maintains a random number generator that generates independent random numbers for different vendors as HMAC salt when performing One-way pseudonym,

$p_{uid\_i} = MD5(salt, uid_i), i = 1, 2, ...n$. This separate mapping ensures that data from different vendors cannot be linked by the pseudonym ID. In addition, the value of the generated p_{uid} is not fixed due to the different random numbers generated, thus achieving the effect of anonymity.

With RBAC, record 3 can be accessed by users who have L1, L2 or L3 privilege.

2. Meanwhile, the host gives credentials to different user groups. The structure of the credentials is shown below. For example, for the credential structure of user group G1:
   - Group: G1
   - Privilege Level: L1
   - Role: governor, police, scientists...
   - Accessible Record: Record 1,2,3

Lastly, the data will be passed from the data host to the proxy host and, after being symmetrically encrypted, to the client.

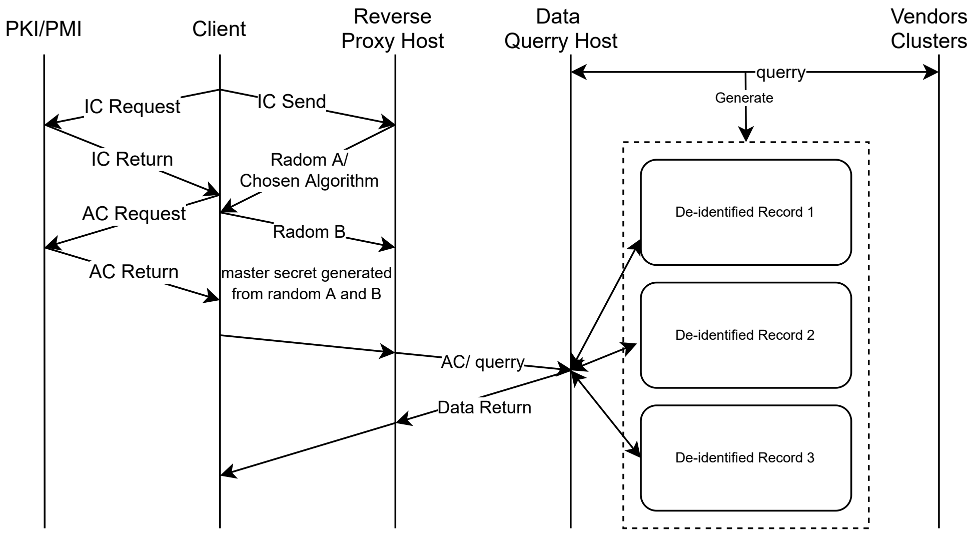

Fig. 1. Messaging diagram

4.2. (b)

We will discuss whether this system satisfies the requirements of the five tasks R1 through R5.

R1. The system supports secure remote three-level privileged access. This can be reflected by RBAC on the data host. It also enables downward access compatibility. As shown above, L1 can access Record 1, 2 and 3 at the same time, while L3 can only access record 3.

R2. The system ensures authorised to access. This can be reflected in the authentication and authorization approach by using ICs and ACs. What's more, With the help of symmetric encryption and the use of message authentication codes, a secure and reliable channel is established between the client and the reverse proxy host. And the bastion host ensures the confidentiality and authenticity of the intranet message transmission by applying the firewall.

Depending on the situation, secure channels may also be established between data hosts and database clusters of different vendors.

R3. The system has minimised key management burdens on both users and vendors, as both client and vendors simply keep their private keys, and maintain a session key in a symmetrically encrypted channel.

R4. The system has minimised communication costs incurred in each data access, as Once the steps of authentication and authorization are completed and during the session lifecycle of the secure channel, no additional handshakes are required between the client and the server, and data requests and transmissions are done in one single pass.

R5. The system has minimised computational costs imposed on the vendors, as All de-identification operations are done in the data host rather than the vendor cluster, and queries to the vendor database are done in a single pass.

4.3. (c)

We will discuss the structure of data request and return messages, as well as discuss the computational cost of the requestor.

1. The data request message will contain the AC, carrying the user role and group information, and get hashed together with the request instruction and transmitted as a whole after being symmetrically encrypted, $Msg_s = AES(MD5(querry, AC), querry, AC, k)||MD5(querry, AC)$. The data reply message is virtually identical to request message, except for the change from querry instucitons to data. Considering that the data may be sliced because of the size, the sequence number seq is introduced, $Msg_r = AES(MD5(data, AC, seq), data, AC, k)||MD5(data, AC, seq)$
2. Before the first request is established, the user's request will additionally include several steps of requesting IC from the PKI, requesting AC from the PMI and key negotiation with the reverse proxy host. When a stable connection is established and during the AC and session lifecycle, the cost to the requestor will be
    1. hashing and encryption of the data.
    2. decryption of the data and hash verification.

4.4. (d)

We will analyze the potential vulnerabilities of this system, and possible ways to improve it.

1. Single point failure and bottleneck. This is reflected in the fact that the bastion host needs to cope with both incoming and outgoing data, while the data host is responsible for all data queries along with the de-identification operations.

To solve this problem, load balancing can be achieved by increasing the number of hosts. In addition, for data hosts, caching of the query data and the mapping between pseudonyms and real names can be achieved by adding Redis hosts to reduce the consumption during repeated queries.

2. There is a flaw in the access control mechanism as the bastion host only requires and verifies the client's IC, leading to the possibility of users who have an IC but are not in user groups G1, 2 and 3 illegally entering the intranet and accessing the data host. Meanwhile, the attacker may also be able to squeeze the bandwidth of normal users by establishing a large number of channels to the bastion host, thus achieving the effect of a dos attack.

   One possible solution is to link the host's access control policy with the firewall policy of the bastion host. Require a legitimate AC before establishing a channel with allocated cache space.

5. References

- Mavridis, I., Georgiadis, C., Pangalos, G., & Khair, M. (2001). Access control based on attribute certificates for medical intranet applications. *Journal of medical Internet research*, *3*(1), E9.
- Linn, J., & Nyström, M. (1999, October). Attribute certification: an enabling technology for delegation and role-based controls in distributed environments. In *Proceedings of the fourth ACM workshop on Role-based access control* (pp. 121-130).
- Noumeir, R., Lemay, A., & Lina, J. M. (2007). Pseudonymization of radiology data for research purposes. *Journal of digital imaging*, *20*(3), 284–295.
- Gergely, A. M., & Crainicu, B. (2016). The concept of a distributed repository for validating X. 509 attribute certificates in a privilege management infrastructure. *Procedia Technology*, *22*, 926-930.
- Chadwick, D. W. (2004). The X. 509 privilege management infrastructure.