Privacy Preserving Multi-Keyword Ranked Search of Medical Data

Beverly Joelyn Fernandes¹, Anusha Pai², Louella Menezes Mesquita e Colaco³
PG Student¹, Assistant Professor²³
Department of Information Technology
Padre Conceicao College of Engineering, Verna, India

Abstract:
Medical records in an electronic format are a uniform way of storing necessary medical data, information and other relevant records about patients with the objective of making easy availability of the necessary data at the required time. These records are of grave importance to the medical establishments, however storing such large amounts of patient records as well as maintaining its privacy is the need of the hour. Utilizing cloud services and encryption techniques will make the needed health care operations more convenient and cost effective. In this paper, we make use of cloud computing techniques that solves the problem of storage, helps in faster retrieval and maintenance of records. This paper also uses hybrid encryption technique (Symmetric and Asymmetric encryption) that ensures privacy and prevents unsolicited access of the medical data thus providing a privacy preserving encrypted cloud data search service. While on the other side to meet the efficient data retrieval process, we need the cloud server to perform result relevance ranking on the large amount of medical records, instead of returning analogous results. We make use of KNN technique to provide necessary relevant data as specified by the search query which is based on multiple keywords. This technique returns records in a ranked order. Ranked search excellently eliminates unwanted network traffic by sending back only the most significant data records.

Keywords: Data privacy, Encryption, RSA, AES, KNN, Ranking, Cloud computing

1. INTRODUCTION

Medical establishments contain Medical records which have patients personal details, disease, forms of examination and instructions made by the respective doctor in series of the undergoing treatment. It helps the patients to achieve the proper and suitable treatment. Furthermore, it is the means for a doctor who look sat the patient. Given the case that each medical establishment has thousands of records to store and maintain, this becomes a tedious task. Thus we need to find an easy and cost effective approach. With the outset of electronic health records, a digital composition of paper records, activities at medical facilities are simplified, more efficient and customer-centric. However by adopting cloud services we can make the health care operations at medical centers more beneficial and profitable. This assures ultimate users to find that cloud computing is the suited choice for their business, since it is less costly than having numerous computers in medical rooms which will need proper hardware, software and network connection for uploading, storing, and retrieving patients or other medical records. With the use of cloud computing, data owners at medical establishments are motivated to outsource their large and complex data management systems from local sites to the profitable public cloud for greater affordability and budgetary savings. Nonetheless for protecting confidentiality of its data, the precise or delicate data needs to be encrypted before deploying, which depletes traditional data utilization established on plaintext keyword search. Hence, enabling an encrypted cloud data search service is of chief importance. Cloud computing is the long wished goal of computing as a service, where cloud clients are allowed to remotely store their records into the cloud so as to enjoy the required quality applications and services from a shared pool of composed computing resources. Its great edibility and budgetary savings motivate both individuals and enterprises to deploy their local complex data management system into the cloud. Protecting data against fraud and unofficial access refers to Cloud data security. To protect human resources and enterprise data in Cloud model is the main target behind data security. For protecting the confidentiality of data and fighting uninvited accesses on the cloud, delicate data for e.g., electronic mails, medical records, photo collections, tax documents, financial agreements etc, will need to be encrypted by data holders before deploying to the profitable public cloud, this, nonetheless, depletes the traditional data application service based on plaintext keyword search. The difficult solution of downloading all the data and decrypting locally is illogical, due to the huge amount of bandwidth cost in cloud scale systems. Thus it is of utmost importance that encrypted cloud data should make use of a Privacy preserving and effective search service. Given that at a particular point of time there are potentially large number of data users in medical establishments and huge amount of outsourced medical records in the cloud, the problem is extremely challenging based on meeting the basic needs of performance, system usability and scalability. On one hand there is a need to meet the effective retrieval process, for that the large amount of data needs result relevance ranking to be performed by the cloud server instead of returning results in an analogous manner. Ranked search using KNN technique helps the data users to find the most relevant information that they need very swiftly and efficiently rather than cumbersomely categorizing through every match in the data assemble. Ranked search can be aided by using multiple keywords in the search request that the user is allowed to specify based on his/her needs. The cloud server performs ranking based on these specified keywords and displays the most relevant data by eliminating unnecessary network traffic which is highly desirable in the “pay as you use” cloud paradigm. On the other hand we have to consider the search result accuracy and strive to enhance and improve the searching based on user experience. It is also necessary for the ranking systems to
support multiple keywords in the search request because single keyword search leads to a lumpy set of results. Just like nowadays web search engines, data users at the medical centers will be able to provide a set of keywords in the search query instead of using single keywords thus giving them most relevant results to their interest. This search using multiple keywords is intended in a manner where in each keyword is used to narrow down the searching results further.

II. RELATED WORK

Cloud computing facilitates a favorable model to data owners for data outsourcing and evades their capital expense by helping them to economically store their data to the public cloud for beneficial data storage and access. Ning Cao et al.[1] defined the problem and came up with a solution for privacy preserving multi keyword search over encrypted cloud data and established many privacy requirements. Among various similarity measures they chose semantics like “coordinate matching” and “inner product similarity” for qualitative evaluation. Also to meet the problem of supporting multi-keyword semantic without any privacy breaches, they proposed a basic idea of MRSE using secure inner product computation and gave two significant MRSE schemes that achieved various stringent privacy requirements in two different threat models. Dawn Xiaodong et al.[3] claimed that storing data on data storage servers such as mail servers and file servers in encrypted form is desirable to reduce security and privacy risks. They described their cryptographic schemes for the problem of searching on encrypted data and provide proofs of security for the resulting crypto systems. Their techniques were provably secure which provided provable secrecy for encryption, such that the untrusted server couldn’t learn anything about the plaintext when only given the ciphertext. Cong Wang et al.[4] were motivated to solve the problem of supporting efficient ranked keyword search for achieving effective utilization of remotely stored encrypted data in Cloud Computing. They worked on a basic scheme and showed that by following an existing searchable encryption framework, it was very inefficient to achieve ranked search. Thus they appropriately weakened the security guarantee and resorted to the newly developed crypto primitive OPSE, and derived an efficient one-to-many order-preserving mapping function, which allows the effective RSSE to be designed. Through thorough security analysis they showed that their proposed solution was secure and privacy-preserving. Cong Wang et al.[2] worked on the crypto primitive OPSE, and derived an efficient one-to-many order preserving mapping function, which allows the effective RSSE to be designed. Their investigation of ranked search mechanism included an efficient support of relevance score dynamics, authentication of ranked search results. Kui Ren et al.[5] studied and defined the problem of fuzzy keyword search over encrypted cloud data for various types of user inputs and the overall user search experience while maintaining keyword privacy.

Hussain Abo Surrah[6] studied data security by including encryption and decryption methods that provided Authorization of the data users directly by the administrators which allowed the files to transfer more securely over the network. Encryption and decryption of both file name and file used asymmetric and symmetric key algorithms respectively. Cong Wang et al.[7] worked on a proposed privacy-preserving public auditing system that was used for data storage security in cloud computing. They used a homomorphic linear authenticator and random masking to guarantee that the TPA would not be able to learn any knowledge about the data content stored on the cloud server during the efficient auditing process, which eliminated the burden of cloud user from the tedious and expensive auditing task and also lightened the user’s fear of the outsourced data leakage. Comprehensive analysis proved that schemes used were probably secure and highly efficient. The experiment that was conducted on Amazon EC2 instance demonstrated the fast performance of their design on both the cloud and the auditor side.

III. SYSTEM ARCHITECTURE

Let us consider a medical data cloud hosting service that includes three main actors namely:
1. The data owner who has a collection of medical records
2. The data user who has the right to search for required data
3. The cloud server responsible for storage and ranking of data

![System Architecture](http://ijesc.org/)

Figure 1. System Architecture

The data owner has a collection of data documents i.e the medical records that is to be sent to the cloud server in the encrypted form. To facilitate searching capacity over encrypted data for effective data usage, the data owner outsources the encrypted document assembly to the cloud server. To search the document collection for k given keywords, a certified user sends a search request containing multiple keywords to the cloud server. The cloud server will perform relevance ranking and send back top-t documents that are most relevant to the search query.

System setup: The user application was created using NetBeans IDE 8.0 and is running over the Amazon cloud server and uses the Relational Database (RDS) for storing patient details.

IV. METHODOLOGY

This system makes use of the following techniques:

A. Hybrid Encryption

Since both asymmetric and symmetric key cryptography have their own pros and cons, modern data transfer systems generally employ a hybrid of the two. Hybrid cryptography use asymmetric keys to initially encrypt symmetric keys known as secret keys. The secret keys are then the ones used to encrypt the actual data. The steps taken by a hybrid cryptography used in secure Data transfer systems are:

1. The cloud server sends its public key to an end user's data transfer client.
2. The client generates a secret key, then encrypts copy of the secret key using the public key of the server and sends that copy to the cloud server.
3. The cloud server receives its copy of the secret key and both of them use that secret key to encrypt/decrypt files exchanged within that period.

![Diagram of hybrid encryption](image)

**Figure 2. Steps in Hybrid Encryption**

Here hybrid encryption technique makes use of:
1. Asymmetric Encryption (RSA) to first generate the keys (public, private and secret key) and then encrypt and decrypt the secret key using the public key and private key respectively.
2. Symmetric Encryption (AES) is used to encrypt the medical records using the secret key.

### B. Asymmetric Encryption (RSA)

Consider a cloud server and a data owner who has to encrypt a file first before uploading it to the server, it would typically be the duty of the server to generate the key pair. It should then send the public key to the data owner and leave the private key on itself.

**RSA Cryptosystem:**

Asymmetric key encryption uses two keys: a private key and a public key. The public key is generated for encrypting the data (in this case it is the secret key), on the other hand the private key is used for decrypting the data (secret key).

**The public-key encryption system has three main phases:**

**I. Key Generation:**

In this project the server needs to receive secret messages so it creates a public key and a private key (which is kept secret). The keys are generated such that they obscure their structure and make it problematic to find the private key just by knowing the public key.

The key generation steps are:

- **Step 1:** select two prime numbers p and q
- **Step 2:** calculate the modulus n = pq
- **Step 3:** calculate the \( \phi(n) = (p-1)(q-1) \)
- **Step 4:** choose a public key exponent integer e such that \( 1 < e < \phi(n) \) and \( \gcd(\phi(n), e) = 1 \)
- **Step 5:** calculate the private key exponent value d such that \( e \cdot d \equiv 1 \mod \phi(n) \)
- **Step 6:** public key = \((e, n)\)
- **Step 7:** private key = \((d, n)\)

**II. Encryption:**

Given the scenario of this particular project the public key is used to encrypt the secret key that is generated by the cloud server. It is formulated as: \( C = M^e \mod n \)

- **Where C = ciphertext**
- **M = message text** (in this case the secret key)
- **e = public key**
- **d = private key**

**III. Decryption:**

The reverse process of encryption will be decryption. Only the cloud server being addressed can easily decrypt the secret key using the private key.

It can be generated using the formula: \( M = c^d \mod n \).

Where \( C = \text{ciphertext} \)

\( M = \text{message text} \) (in this case the secret key)

\( e = \text{public key} \)

\( d = \text{private key} \)

### C. Symmetric Encryption (AES)

AES is a symmetric block cipher that was published in 2001 and intended to replace DES and 3DES. It is made up of 128/192/256 bit keys and 128 bit data. It processes data as block of 4 columns of 4 bytes and operates on entire data block in every round. In this project we make use of AES to encrypt and decrypt medical records using the secret key generated by the cloud server.

**Plaintext:** The original message and the one that is to be encrypted.

**Ciphertext:** The encrypted message which is the output of the algorithm.

**Block Size:** The number of bits that the cipher will operate on. This is 128 bits in AES. Each message is split up into this number of bits and the operations described below carried out.

The Advanced Encryption Standard has the following main stages:

1. **Key Expansion - Round keys are derived from the cipher key using Rijndael’s key schedule.**
2. **The Initial Round - This is a straightforward step, take the plaintext and XOR it with the encryption key.**
3. **The Intermediate round - This stage has four steps:**
   - a) **SubBytes** – use the s-box to replace each byte of the state with a new value.
   - b) **ShiftRows** – Rotate the state in a prescribed fashion.
   - c) **MixColumns** – Take 4 bytes and mix them so that all four bytes have an effect on the value of each of the resulting four bytes.
   - d) **AddRoundKey** – each byte of the state is combined with the round key created by the key schedule above.
4. **The Final Round - There are more attacks being devised against AES and as computing capacity improves, developments of these attacks may become practicable”**

**This round carries out three steps:**

1. **SubBytes**
2. **ShiftRows**
3. **AddRoundKey**

The reason why the final round does not have a ‘mixcolumns’ step is because that step is used to feed into the next round. Since this is the final step and there is no next round, the final round excludes that step. Finally the state is encrypted.

### D. KNN and Ranking

The k-Nearest Neighbors algorithm (or KNN for short) is a non-parametric method used for classification, the input consists of the k closest training examples in the feature space.[8]. In the secure k-nearest neighbor (KNN) scheme, Euclidean distance between a database record p and a query vector q is used to select the k nearest database records.

**Euclidean Distance** = \( \sqrt{(x^2 - y^2)} \)
KNN aims to get the closest cluster of records to the search query containing multiple keywords. A search query is classified by a majority vote of its neighbors, with the query being assigned to the class most common among its k nearest neighbors where k is a positive integer and is typically small.

The system consists of two panels. One the doctor panel and other the search panel. The former is used for entering medical records while the latter is used for searching stored medical records from the database. The search on medical records is based on diagnosis, symptoms and medication. The search query will contain keywords related to the above three categories. An Index is constructed for each of the categories which is used for searching. The search query is also converted to an index (numerical form). The search is done using KNN technique where the smallest possible difference is found between the search query and the stored data records. The search query is compared with the index of database entries from top to down. The smallest difference is then ranked as the topmost search result followed by the results with greater difference.

E. Amazon Cloud
With growing demand for physical space, elasticity demand, cost savings and speed, cloud based services act as an ideal remedy to large businesses and organizations. Amazon Web Services offers reliable, scalable, and inexpensive cloud computing services.

The proposed system makes use of the following Amazon Cloud Services:

1. AWS Elastic Beanstalk
Elastic Beanstalk is called a Platform As A Service (PaaS) since it allows users to create, push and manage web applications in the Amazon Web Services console. With Elastic Beanstalk, we can quickly deploy and manage applications in the AWS Cloud without worrying about the infrastructure that runs those applications. AWS Elastic Beanstalk helps reduce management complexity without restricting choice or control. Once your application is uploaded, Elastic Beanstalk consequently handles the details of space arrangement, load balancing, scaling, and application health monitoring. It uses highly dependable and expandable services that are available in the AWS Free Usage Tier.

2. RDS
Using Amazon Relational Database Service (Amazon RDS) makes it easier to set up, operate, and expand relational databases in the cloud. It provides cost-efficient and sizable capacity while managing time-consuming database administration tasks that enables one to focus on its applications and business.

3. IAM
AWS Identity and Access Management (IAM) is a web service that helps you protect control access to AWS resources for your users. You use IAM to control that can use your AWS resources and what resources they can use and in what ways. In other words it supports authorization and authentication.

V. RESULTS
The system consists of the following 3 panels:

1. The System Login Panel
Allows access to only authorized users of the system.

2. The Doctor Panel
This panel helps doctors to enter the necessary patient details into the system.

3. The Intern Panel
This panel is used by an intern or any authorized data user to search for required medical records.

The following are the results for search by disease:

Search query - Acute appendicitis
The first result returned is the most relevant to the search query and is first entry found while scanning the database using knn technique which is then followed by the next relevant result that is returned for that particular search query.

The following are the results for search by symptoms:
Search query - abdominal pain

Figure 3. The System Login Panel

Figure 4. The Doctor Panel for Entering Medical Records

Figure 5. The Intern Or User Panel For Searching Medical Records

Figure 6. Results for Search by Disease

Figure 7. Results For Search By Disease
VI. CONCLUSION

In this paper we make use of asymmetric (RSA) and symmetric encryption (AES) techniques which protects the privacy of medical records over a cloud server. The use of a cloud system solves the problem of storage for large amount of medical records in medical establishments and also simplifies its accessibility. The search of the medical records is carried out using KNN technique wherein the search query consists of multiple keywords and the results are displayed in a ranked order giving the most relevant results to the users.

VII. REFERENCES


[8]. www.wikipedia.org

[9]. www.aws.amazon.com