Multi Keyword Similarity Search Over Encrypted Text Data on Cloud

Sumeet Pinjarkar, Tushar Saindane, Pradip Shinde, Payal Rahatal, Amol Kekan,
Student, D. Y. Patil College of Engineering, Akurdi, Pune
Mrs. N. S. Patil Assistant Professor,
D. Y. Patil College of Engineering, Akurdi, Pune.

Abstract - The tremendous amount of data is being outsourced every day by individuals or enterprises . It is not feasible to manage or to store such a large data locally, due to the limited storage capacities, and the system becomes the single point of failure. the cloud comes into picture to store the data with better flexibility and cost saving. As the data might be confidential or sensitive, the data which user wants to store on the cloud can be private and it should not be leaked, for that purpose searchable encryption is be used, so that even if the file falls in wrong hands it will be safe. At the time of retrieval of data, the multi-keyword search over text data can only handle the exact keywork matching. Multi-keyword similarity search overcomes the problem of not finding any related documents on searching. while encrypting the data before storing it to the cloud will help to preserve the privacy of the files. Finding the similarities between input keyword or similar keyword is done by edit distance metric algorithm. Final design to achieve the user privacy, and to speedup the search task. At cloud side Bloom Inverted List is used to implement searching on index.

Keywords—*Software as a Service, Platform as a Service,Infrastructure as a Service, Bloom Filter, Bloom Inverted List.*

____*****____

I. INTRODUCTION

Every individual is producing tremendous amount of date than ever before, and this rate is only going to increase dayby-day. Also more importantly the organizations have much higher rate of producing data which is in fact more sensitive too. Hence, organizations are often more concerned about the security of their data to store it on cloud storage, all of this leads to the increased authentication demand. Considering the privacy of the data over the cloud, the searching techniques should be good enough to not to expose the data publicly, it will require us to encrypt the data on cloud. Searching is not feasible to do with traditional encryption schemes. Enhanced and more sophisticated cryptography may offer new tools to make the data searchable encrypted. Encryption schemes like searchable encryption also known as predicate encryption that allow operation and computation on the cyphertext, allows the data owner to compute a capability from his secret key. A capability encodes a search query, and the cloud can use this capability to decide which documents match the search query, without the requirement of any additional information. Other cryptographic techniques such as homomorphic encryption and Private Information Retrieval (PIR) perform computations on encrypted data without decrypting it

A. Searchable encryption:

Searchable encryption allows a party to outsource the data in a private manner, while maintaining the ability to selectively search over it. In the process of searching on private-keyencrypted data, the user himself encrypts the data, so he can organize it in an arbitrary way before encryption and include additional data structures to allow for efficient access of relevant data. The data and the additional data structures can then be encrypted and stored on the server so that only someone with the private key can access it. In this setting, the initial work for the user is at least as large as the data, but subsequent work of accessing the Data is very small relative to the size of the data for both the user and the server[3].

B. Bloom Inverted List:

Bloom Inverted List is the data structure which is constructed by combination of Inverted list and Bloom Filter data structure[4].

1) Inverted List:

Inverted List is a fundamental technique to support keyword search. It can be considered as a 2-D array, each row of which consists of a tuple. The length of the array is equal to number of keywords. And First entry of each tuple is a keyword and second entry of tuple consists of file ID. This Inverted List can be encrypted to form an Encrypted Inserted List[5].

2) Bloom Filter:

A Bloom Filter is a simple space-efficient, probabilistic data structure for representing a set in order to support membership testing queries. The first column of the Bloom Inverted List corresponds to the hashes of the first column of the Inverted List. File Ids in the Inverted List are all stored in second column in the form of Bloom Filter. User calculate and submit bloom filter of keywords to cloud. In Bloom Inverted List each tuple has file ids as its first column and the bloom filter of keywords as the second column[5].

C. *Amazon EC2*

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is make web-scale cloud computing easier for developers.

Amazon EC2 provides web service interface which allows you to obtain and configure capacity with ease. It provides you with complete control of your computing resources and lets you run on Amazons proven computing environment.It provides quick process of starting and managing the server. Amazon EC2 is economical as you have to pay only for capacity that you actually use.

D. *Bitnami*

Bitnami makes it easy to deploy apps with native installers like XAMP, as virtual machines, or in the cloud. Cloud servers provided by google cloud service or Amazon cloud service or Oracle cloud etc can be connected with Bitnami and the apps can be hosted throuhg Bitnami on these servers

II. LITERATURE REVIEW

A. *Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data.*

the words which are entered for querying are searchable. When you type these words into the database search window, this is called keyword searching. In Multi-Keyword searching all of the files containing at least threshold number of keywords specified by the user will be returned to the user.

B. *Similarity Search:*

Nearest neighbour search and Range queries are important subclasses of similarity search, and a number of solutions exist. Research in Similarity Search is dominated by the inherent problems of searching over complex objects[4].

C. *Nearest neighbour search (NNS) :*

Also known as proximity search, similarity search, is an optimization problem for finding closest (or most similar) points. Closeness is typically expressed in terms of a dissimilarity function i.e. the less similar the objects, the larger the function values.

III. SYSTEM ARCHITECTURE

Searching architecture for encrypted cloud data includes following components: Data Owner, Data User, Cloud Storage or Server.

A. *Data Owner*

Data owner extract keyword from data collection. He construct searchable encrypted index from the data i.e Bloom Inverted List, then he encrypt file and send both index and encrypted file to cloud server.

B. *Data User*

User requests in the form of keywords to the cloud server.

C. *Cloud Server*

Receives the request from user and then send the corresponding encrypted files to the user as response.

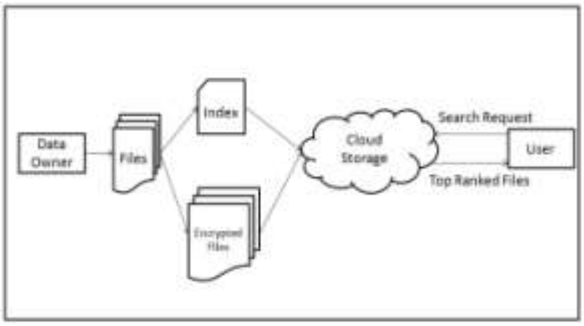

Fig. 1. Cloud Computing System

IV. PROPOSED SYSTEM

This system combine the idea about the encrypted inverted list and the idea about the Bloom table to construct a new data structure to support the MKSS. The owner first constructs the inverted list. Instead of constructing the corresponding encrypted inverted list, the owner constructs a so-called Bloom inverted list in pre-processing phase.

Algorithm:

Preprocessing Phase:

1 Firstly owner creates the Bloom Inverted List I i.e. B(Sw,d).

2 After suppressing, owner computes hashes on both columns in BIL, using two secrete keys and then send it to the cloud.

Searching Phase:

1 User send the keywords to the cloud. i.e. B(Sw,d). 2 Owner calculates all possible keywords and send to Bloom FIlter.

3 Bloom Filter check membership of keyword in the files. 4 Cloud construct set of file IDs for each keyword.

5 Then IDs of files in the intersection set of the list of sets from above step are used to select the files and files are then decrypted.

6 User retrieve the decrypted files.

V. EXPECTED RESULT

User can upload file and it will be stored in encrypted form. Searching will not expose any data to data owner unless he knows how to decrypt it. Bloom filter enables quick membership testing on files.

REFERENCES

[1] Hanhua Chen , Hai Jin , Lei Chen, Yunhao Liu, and Lionel M. Ni.,*Optimizing Bloom Filter Settings in Peer-to-Peer Multikeyword Searching.* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.24, NO. 4, APRIL 2012.

[2] Ning Cao, Cong Wang, Ming Li, Kui Ren, and Wenjing Lou.,Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data.IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25,NO. 1, JANUARY 2014

[3] Muhammad Naveed, Manoj Prabhakaran, Carl A. Gunter University of Illinois at Urbana-Champaign. Dynamic Searchable Encryption via Blind Storage. 2014 IEEE Symposium on Security and Privacy.

[4] Chia-Mu Yu, Chi-Yuan Chen, and Han-Chieh Chao, Senior Member, IEEEPrivacy-Preserving Multikeyword Similarity Search Over Outsourced Cloud Data2015 IEEE

[5] Sasu Tarkoma, Christian Esteve Rothenberg, and Eemil Lagerspetz Theory and Practice *of Bloom Filters for Distributed Systems*