ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 399 - 402 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Privacy Preserving and Optimal Keyword Search Over Out Sourced Data

[1]J. Satyanarayana Swamy, [2]V. Visweswararao
[1]Final M.Tech Student, [2]Assistant Professor
[1,2]Dept of Computer Science and Engineering
[1,2]PYDAH College of Engineering, Visakhapatnam, AP, India

Abstract: Searching top k multi keywords from the out sourced data files is still an interesting research issue because out sourced data over cloud can be encrypted for confidentiality .In this paper we are proposing an efficient top k retrieval from out sourced file through service oriented application by computing the file relevance score for input multi keywords and symmetric key encryption and every manipulation comes from the server end instead of client end every the ranking of the documents based on file relevance scores.

INTRODUCTION

Cloud Computing [1] refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS).The datacenter hardware and software which is collectively known SaaS Cloud. When a Cloud is set to get provide in a pay-as-you-go way provided to the common public, we refer it a Public Cloud; the service which is sold known as Utility Computing. The term Private Cloud is to refer for internal datacenters of a commercial or organization purpose not made available to the general public. Thus, Cloud Computing is a combination of Utility Computing and SaaS. We focus on Cloud Computing effects on Cloud Providers and SaaS Providers/Cloud. The top level can be recursive, in that SaaS providers can also be SaaS users via mash-ups.

Private Clouds. People can be users or providers of SaaS, or users or providers of Utility Computing. We focus on SaaS Providers (Cloud Users) and Cloud Providers, which have received less attention than Saas Users. It makes provider-user relationships clear. from an electrical machinery point of view, there are three aspects in Cloud Computing.

1. The illusion of infinite computing resources available on demand, thereby eliminating the need for Cloud Computing users to plan far ahead for provisioning.

2. The elimination of an up-front commitment by Cloud users, thereby allowing companies to start small and increase hardware resources only when there is an increase in their needs.

3. The ability to pay for use of computing resources on a short term basis as needed (e.g., processors by the hour and storage by the day) and release them as needed, thereby rewarding conservation by letting machines and storage go when they are no longer useful.

We argue that the construction and operation of extremely large-scale, commodity-computer datacenters at low-cost locations was the key necessary enabler of Cloud Computing, for they uncovered the factors of 5 to 7 decrease in cost of electricity, network bandwidth, operations, software, and hardware available at these very large economies of scale[2][3].

RELATED WORK

Even though various approaches available for searching out sourced information they are not optimal to search multiple queries securely because homomorphism encryption techniques increases the time complexity and increases the size of the cipher text and regular architectures simply work client server architecture with respect to file relevance scores, they are not compatible all sort of applications like standalone web application or mobile application in future [4].

Homomorphism encryption [5]is an encryption technique which is used to convert plain text by specific type of computation into cipher text and the result is decrypted to get the same plain text by converting the cipher text.

This is one such important feature in architecture of modern communication system. Homomorphism encryption without exposing data from services it allow different. services from different commercial organizations such as 1)calculation of tax 2)rate for the currency exchange and 3)shipping, service for the unencrypted data without exposing the data over the service from different services.[9,10] Homomorphism encryption schemes are malleable by design. The homomorphism property of various cryptosystems can be used to create secure voting systems [6, 7, 8] collision-resistant hash functions, private information retrieval schemes and enable widespread use of cloud computing by ensuring the confidentiality of processed data.

There are several efficient, partially homomorphism cryptosystems, and a number of fully homomorphism, but less efficient cryptosystems. Although a cryptosystem which is unintentionally homomorphism can be subject to attacks on this basis, if treated carefully homomorphism can also be used to perform computations securely.

PROPOSED WORK

In this paper we are proposing an efficient top k results approach for the input of multiple keywords forwarded by the user. Initially data owner generates an index table by segmenting the file collection and preprocess the file collection and encrypts the feature information then computes Term frequency and inverse document frequency for the respective files in the file collection for computing file relevance score .Business logic can be maintained in service oriented application, so it provides a common platform either to the web application or mobile application in future if required .user query can be compared with the encrypted word from server end instead of computing everything at client end.

MODULES

BASE TABLE GENERATION

In this approach data owner out sources the data in the server, before storing data in the server , Data owner has a collection of n data files C = (F1, F2, Fn) that he wants to outsource on the server in encrypted form while still keeping the capability to search through them for effective data utilization reasons. To do so, before outsourcing, data owner will first build a secure searchable index I from a set of m distinct keywords W = (w1;w2; :::;wm) extracted from the file collection C, and store both the index I and the encrypted file collection C on the server. After searching the information data can be organized after the ranking.

To do so, before outsourcing, data owner will first build a secure searchable index I from a set of m distinct keywords W = (w1,w2,… ,m) extracted from the file collection Cinder table contains the unique keywords from the datasets along with file ids, before placing them into the index table encrypt the keywords by using symmetric key approach with AES algorithm for security purpose.

Algorithm for Base table generation

1. Read the document F

2. Segment the document term wise and encrypt with key

3. Calculate term frequency (TF) and inverse document frequency (IDF) and publishing time (P_T)

4. Generate index table (I_{table}) and files upload to server

Data owner can store data in cloud server but does not know where data actually stored, so to maintain confidentiality, optimality and security, data can be initially preprocessed by eliminating the unnecessary information from the data component and extracts keyword by keyword, encrypts keyword by using triple DES algorithm and measures term frequency of the keyword, Base table can be generated with three attributes along with file id and uploads to the server. The following table shows sample base table

Keyword	**Cipher Keyword**	**Term Frequency**	**File ID**
Mobile	$%^&*(	4	Abc.html
Apple	*(!!~*^%	3	Hello.docx
Elephant	##‖%^$%&	1	Hello.docx
Paper	$%^$%^	2	Main.txt

Fig1: Base Index Table

KEYWORD RANKING

When a user forwards an input query to service provider, finally rank oriented results can be forwarded to end user based on the term frequency (TF), it computes number of occurrences of a keyword in a document and inverse document frequency (IDF) refers number of occurrences of a keyword in all the documents. File relevance score can be calculated for every individual document in terms of TF and IDF and forward File relevance score based descending order of results to input query.

F_{score}= TF * IDF

Here we need to calculate File relevance score with respect to multiple keywords of input query, so we need to order file relevance score with respect to

Web service

Web services are service oriented application which can maintains business logic in centralized location instead of maintaining processing module at every individual client end or at user interface, it reduces the redundancy, minimizes the malfunction chances by maintaining the business logic way from the end users. The main advantage of the web services are language interoperability, any standard programming language can communicate with web service native language by using intermediate language web service description Language.

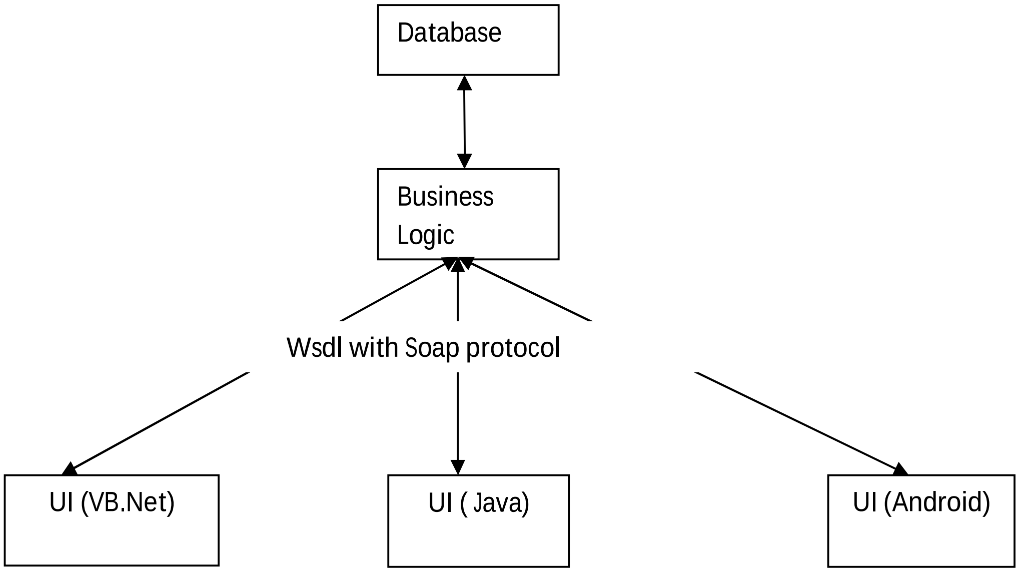

Fig 2 : Web service

web service and client (User interface) can communicate with each other in terms of web service description language (wsdl),every object can be transmitted through soap protocol (i.e everything is a soap object)

SEARCH IMPLEMENTATION

End user initially registers at data owner to receive key which is used to decrypt the cipher text in base table, after login end user can be authenticated with user id and key and whenever a user forwards an input query, it can be converted to cipher query and compares with cipher keywords in base table and retrieves TF and IDF of cipher keyword and computes file relevance score, instead of comparing plain textual information.

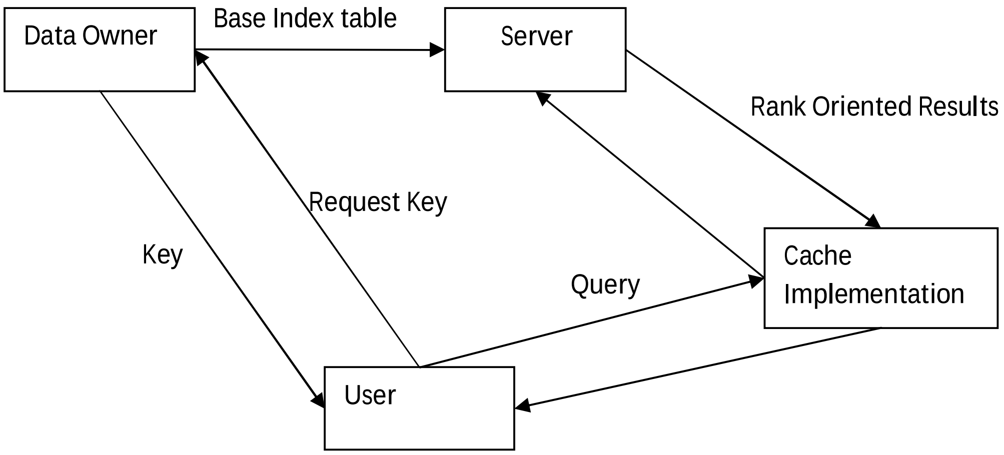

Fig 3: Proposed Architecture

Files can be retrieved based on the our novel file relevance scores

Step1: Registration of the user at Server by requesting the key

Step2: User receives the key for authenticated and secure search

Step3: User searches for relevant data with a plain keyword

Step4: Service process the query and checks for the authentication of user

Step5: Service retrieves the relevant information from index table for respective keyword

Step6: calculates the file relevance scores based on the file relevance score

relevance_Scores[j] = Convert.ToDecimal((1 / termsinfile[j]) * (1 + Math.Log(termfreqs[j])) * Math.Log(1 + (filecount / numberoffiles)));

Step7: return the files based on the file relevance score to user

End user request for the key from data owner, Data owner forwards a search key to requested user. User can login with his/ her user-id and key and can search by passing input query, input request forwarded to web service, it initially validates the key, if he is authenticated, keyword can be encrypted and compares with the encrypted keywords in the base tables and computes term frequency and inverse document frequency for file relevance score, returns results based on the score of the files.

CONCLUSION

We are concluding our research with efficient multi keyword search ranking over service oriented applications by computing file relevance score in terms of term frequency and inverse document frequency, our experimental results shows secure and efficient search implementation and language interoperability

REFERENCES

1. ARMBRUST, M., AND ET AL. Above the clouds: A berkeley viewof cloud computing. Tech. Rep. UCB/EECS-2009-28, EECS Department,U.C. Berkeley, Feb 2009.
[2] M. Arrington, "Gmail disaster: Reports of mass email deletions,"http://www.techcrunch.com/2006/12/28/gmail-disasterreports-of-mass-email-deletions/,December 2006.
[3] Amazon.com, "Amazon s3 availability event: July 20, 2008,"http://status.aws.amazon.com/s3-20080720.html, 2008.

[4] RAWA News, "Massive information leak shakes Washington over Afghan war,"http://www.rawa.org/temp/runews/2010/08/20/massive-information-leak-shakeswashington-over-afghan-war.html, 2010

[5] AHN, "Romney hits Obama for security information leakage,"http://gantdaily.com/2012/07/25/romney-hits- obama-for-security-information-leakage/,2012

[6] Cloud Security Alliance, "Top threats to cloud computing," http://www.cloudsecurity alliance.org, 2010.

[7] C. Leslie, "NSA has massive database of Americans' phone calls,"http://usatoday30.usatoday.com/news/washington/2006-05-10/.

[8] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption improved definitions and efficient constructions," in Proc. of ACM CCS, 2006.

[9] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encryptedcloud data," in Proc. of ICDCS, 2010.

[10] S. Zerr, D. Olmedilla, W. Nejdl, and W. Siberski, "Zerber+r: Top-k retrieval from a confidential index," in Proc. of EDBT, 2009.