Towards Differential Query Services in Cost-Efficient Clouds

D.Nagarajan[#1], J.Ambika[*2]
Master of Technology in Computer Science & Engineering, PRIST University
Thanjavur - 613 403.
[1]nagarajaneb78@gmail.com
[*]Assistant Professor, Department of CSE
Kings College of Engineering
[2]ambi.nagu08@gmail.com

***Abstract*— Cloud computing as an emerging technology is expected to reshape information technology processes in the near future. The basic idea of EIRQ is to construct a privacy preserving mask matrix that allows the cloud to filter out a certain percentage of matched files before returning to the ADL. Cloud computing provide cost efficient solution for efficient searching and also provide query services that allows each user retrieve the matched files and using mechanism to retrieval of efficient query services over encrypted data.**

***Keywords*— Cloud Computing, Differential Query services, Computation, Information Retrieval, Ranked Query**

I. INTRODUCTION

Cloud computing as an emerging technology is expected to reshape information technology processes in the near future. Due to the overwhelming merits of cloud computing, e.g., cost-effectivenes, flexibility and scalability, more and more organizations choose to outsource their data for sharing in the cloud. As a typical cloud application, an organization subscribes the cloud services and authorizes its staff to share files in the cloud. Each file is described by a set of keywords, and the staff, as authorized users, can retrieve files of their interests by querying the cloud with certain keywords. In such an environment, how to protect user privacy from the cloud, which is a third party outside the security boundary of the organization, becomes a key problem.

II. RELATED WORK

Private searching was proposed by Ostrovsky allows user to retrieve files of interest from an untrusted server without leaking any information. Otherwise, the cloud will learn that certain files, without processing, are of no interest to the user. Commercial clouds follow a pay-as-you-go model [8], where the customer is billed for different operations such as bandwidth, CPU time, and so on. To make private searching applicable in a cloud environment, our previous work designed a cooperate private searching protocol (COPS) [6], where a proxy server, called the aggregation and distribution layer (ADL), is introduced between the users and the cloud. Our aim is to protect user privacy through differential query services by Aggregation and distributions Layer. The ADL deployed inside an organization has two main functionalities: aggregating user queries and distributing search results. Under the ADL, the computation cost incurred on the cloud can be largely reduced, since the cloud only needs to execute a combined query once, no matter how many users are executing queries. Furthermore, the communication cost incurred on the cloud will also be reduced, since files shared by the users need to be returned only once. Most importantly, by using a series of secure functions, COPS can protect user privacy from the ADL, the cloud, and other users. The problems with existing scheme have a high computational cost [2], since it requires the cloud to process the query on every file in a collection. It will quickly become a performance bottleneck when the cloud needs to process thousands of queries over a collection of hundreds of thousands of files [1]. That is the reason we shift our momentum of research towards differential query services with ADL in order to reduce computational cost, low bandwidth usage.

III.USER PRIVACY

User privacy can be classified into search privacy and access privacy. Search privacy means that the cloud knows nothing about what the user is searching for, and access privacy means that the cloud knows nothing about which files are returned to the user. When the files are stored in the clear forms, a naive solution to protect user privacy is for the user to request all of the files from the cloud; this way, the cloud cannot know which files the user is really interested in. While this does provide the necessary privacy, the communication cost is high.

A. Private Searching

Private searching was proposed by Ostrovsky et al, which allows a user to retrieve files of interest from an untrusted server without leaking any information. However, the Ostrovsky scheme has a high computational cost, since it requires the cloud to process the query (perform homomorphic encryption) on every file in a collection.

Otherwise, the cloud will learn that certain files, without processing, are of no interest to the user. It will quickly become a performance bottleneck when the cloud needs to process thousands of queries over a collection of hundreds of thousands of files. We argue that subsequently proposed improvements, like also have the same drawback. Commercial clouds follow a pay-as-you-go model, where the customer is billed for different operations such as bandwidth, CPU time, and so on. Solutions that incur excessive computation and communication costs are unacceptable to customers.

B. ADL - Proxy Server

To make private searching applicable in a cloud environment, our previous work designed a cooperate private searching protocol (COPS), where a proxy server, called the aggregation and distribution layer (ADL), is introduced between the users and the cloud. The ADL deployed inside an organization has two main functionalities: aggregating user queries and distributing search results. Under the ADL, the computation cost incurred on the cloud can be largely reduced, since the cloud only needs to execute a combined query once, no matter how many users are executing queries. Furthermore, the communication cost incurred on the cloud will also be reduced, since files shared by the users need to be returned only once. Most importantly, by using a series of secure functions, COPS can protect user privacy from the ADL, the cloud, and other users.

C. Differential Query Services

A novel concept, differential query services, to COPS, where the users are allowed to personally decide how many matched files will be returned. This is motivated by the fact that under certain cases, there are a lot of files matching a user's query, but the user is interested in only a certain percentage of matched files. To illustrate, let us assume that Alice wants to retrieve 2% of the files that contain keywords "A, B", and Bob wants to retrieve 20% of the files thatcontain keywords "A, C". The cloud holds 1,000 files,where {F1,...,F500} and {F501,...,F1000} are describedby keywords "A, B" and "A, C", respectively. In theOstrovsky scheme, the cloud will have to return 2, 000 files. In the COPS scheme, the cloud will have to return 1, 000 files. In our scheme, the cloud only needs to return 200 files. Therefore, by allowing the users to retrieve matched files on demand, the bandwidth consumed in the cloud can be largely reduced.

D.EIRQ

A scheme, termed Efficient Information retrieval for Ranked Query (EIRQ), in which each user can choose the rank of his query to determine the percentage of matched files to be returned. The basic idea of EIRQ is to construct a privacy preserving mask matrix that allows the cloud to filter out a certain percentage of matched files before returning to the ADL. This is not a trivial work, since the cloud needs to correctly filter out files according to the rank of queries without knowing anything about user privacy.

IV. MODULES

Focusing on different design goals, we provide two extensions: the first extension emphasizes simplicity by requiring the least amount of modifications from the Ostrovsky scheme, and the second extension emphasizes privacy by leaking the least amount of information to the cloud.

A. Login credential in cloud:

Normally the cloud is used to perform reduce the cost of data warehouse infrastructure for a corporate network companies. The corporate company databases are stored in cloud for instead of creation of warehouse management on that company. So the company details and database store to cloud with high security credential information such as id and password. The id and password is varying it each individual workers of a company. So if the user login credential match means to entering a cloud computing network.

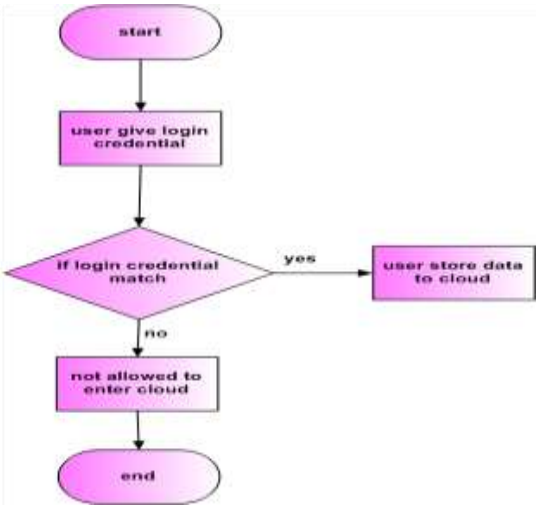

Fig 1 Login **Credential**

B. *File retrieval using cloud*

The purpose of cloud is the data's are retrieved by the big companies in a convenient manner in cost based query services. Here the user login credential is match means the user upload the data to a cloud in a secure manner.

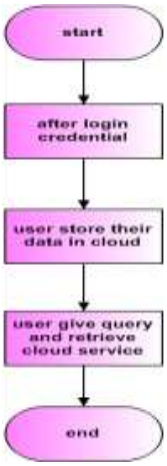

Fig 2 File Retrieval

The user want to retrieve data means provide the query to get relevant content of user search base queries.

C. *Minimizing query overhead on cloud*

The user stores their data in a cloud on secure manner. The one user from company to give the query and get result in query base services in a cloud in proper secure way. But this

is suitable for a single user. Suppose multi user given query to a same cloud means the cloud server of process time is high. So the cost of the query services in cloud is high. So our module is to reduce the query overhead.

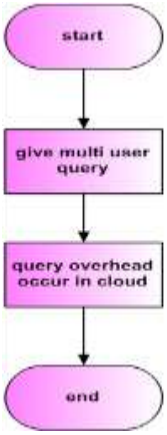

Fig 3 Minimizing Query Overhead

D. Perform ranking accuracy in cloud

If the multi user given query to the same cloud, so the processing time is high. So the priority order of users queries process to provide efficient query services on the cloud. For instance we consider 100 employee in a company means the 1st 30 members are search a same query in cloud, next 15 members search another query keyword and finally last 55 members are search another query keyword.

Fig 4 Ranking in Cloud

The cloud performs priority based which query is mostly given by the users and calculating the ranking of query and aggregate the users are based on keyword query. So these ways of (EIRP) approach to reduce the processing time in cloud and reduce query overhead in cloud.

E) Architecture

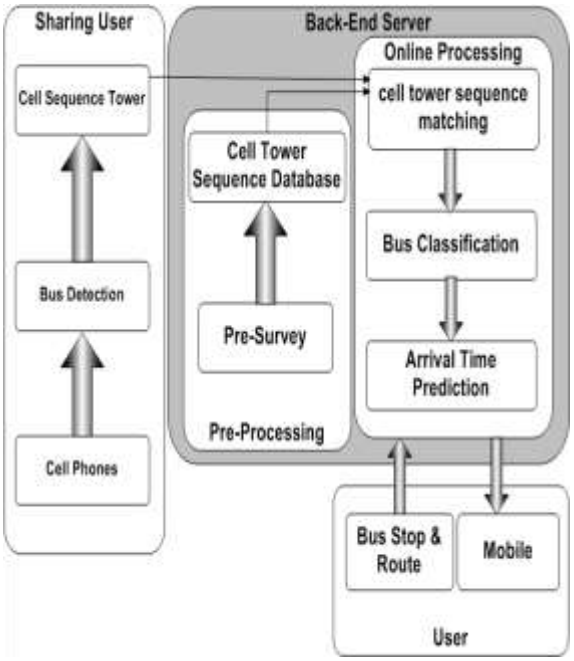

Fig 5 System Architecture

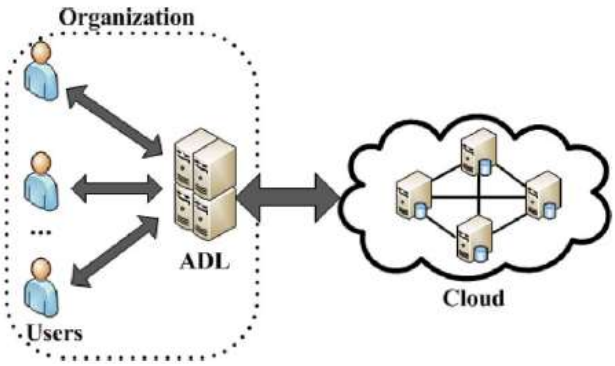

Fig 6 Process Flow

F) Implementation of EIRQ

EIRQ consists of four algorithms; the working process is shown in Fig 6.

1. The user runs the GenQuery algorithm to send rank of the query and keywords to the ADL. The query can be sent without encryption, with the help of ADL.

2. After combining the user queries, the ADL runs the ConstructMatrix algorithm to send maskmatrix to the cloud. The maskmatrix M consists of d-row and r-column matrix, where d denotes the number of keywords in the dictionary, and r denotes the lowest query rank.
   Let M[a,g] denote the element in the a-th row and j-th column and l be the highest rank of queries that choose a-th keyword in the Dic[a] in the dictionary. Maskmatrix M is constructed as follows: for a-th row of M that corresponds to Dic[a], M[a,l],…,M[a,r-l] are

set to 1, and M[a,r-l+1],….,M[a,r] are set to 0. Rather than choosing the random r-l elements, the ADL sets the first r-l element to 1. The probability of k-th element that corresponds Fj keywords being 0 is l/r.

3. The cloud runs the FileFilter algorithm to return the buffer. The buffer also contains the matched files to the ADL. Corresponds to Fjkeyword the cloud multiplies the k-th element to form cj , where k= g mod r. In ostrovsky scheme, c-e pairs into the multiple entries of the buffer. From the step, Rank-l file Fj, the probability of cjbeing 0 is l/r, thus the probability of Fjbeing filtered out is l/r.
4. The ADL runs the DistributeResult algorithm to distribute the search results to each user. The cloud attaches the keywords to the file contents, so the ADL is able to distribute files correctly. By executing user queries, the ADL can find all matched of the users queries.

V ANALYSIS

Here we are proposing two way of analysis. First one is Security analysis and second one is performance analysis. 6.1

1) Secure analysis

In security analysis of EIRQ-Scheme we can provide search privacy, access privacy and rank privacy. In search privacy scheme the combined query sent to the cloud is encrypted under the ADL's public key with the Paillier cryptosystem. The paillier cryptosystem is semantically secure and the cipher text of every 1 or 0's . So the cloud can't deduce what each user is searching from the encrypted query. In access privacy scheme the cloud processes the encrypted query on each file in a collection and maps the processing result into the buffer, which is encrypted by the ADL's public key. So that the cloud cannot know which file is returned from encrypted buffer. In rank privacy scheme the message from the ADL to the cloud are r-encrypted query, the buffer size and the mapping time. In EIRQ-Privacy, the message from the ADL to the cloud is a d-row and m column mask matrix, where d is the number of keywords in the dictionary, and m is the maximal value of mapping times.

2) Performance analysis

We are comparing the performance between no rank and the three EIRQ schemes under different parameters .In no rank the ADL only combines the user queries but it will not provide user differential queries. In the supplementary file we also provide the comparison between the no rank and the work performance Suppose that queries are classified into 0 ∼ r ranks, where t files stored in the cloud whose keywords constitute a dictionary of size d, fi files matching Rank-i queries, and f_i files matching Rank-i queries but mismatching higher ranked queries. Furthermore, in No Rank and EIRQ-Efficient, the threshold file survival rate p_ is set to α; in EIRQ-Simple and EIRQ-Privacy, p_ i is set to i/r+α.

3) File Survival Analysis

Now a days query are divided into 0 ∼ 4 rank, queries in Rank-0, Rank-1, Rank-2, Rank-3, and Rank-4 would retrieve 100%, 75%, 50%, 25%, 0% of matched file, individually. However, the real failure rates in a EIRQ-Simple and EIRQPrivacy between the Ostrovsky parameters setting is much less than i/r, and so, the real file survival rates is greater than the desired value of 1 − i/r (about 25% and 50% of files are redundantly returned to users); Only EIRQ-Efficient, which filters a certain percentage of a matched file previous mapping them to the buffer, provides distinguishable query service.Under a Bloom filter parameter settings, first we obtain corresponding mapping time. Significantly, for the file survival rate 100%, 75%, 50%, 25%, we have the alternate mapping times 7, 2, 1, 0.4, resp. Based on this values, the buffer size can be calculate with Eqs. 4-6 for 5 10 15 20 25 50 .

4) Communication Analysis

The communication cost mainly depends on the buffer size generated by the cloud, which is calculated in different ways under different parameter settings. Furthermore, the buffer size depends on the number of files that match the queries, which is different when users have different common interests, i.e., the average number of common keywords among user queries. Therefore, in different parameter settings, we will analyze the buffer size under different common interests. In the following experiments, 1 common keyword, 2 common keywords, and 4 common keywords denote that the average common keywords among user queries are 1, 2, and 4, respectively; random keywords denote that each user randomly chooses keywords for its query.

VI Conclusions

In this paper, efficient information retrieval for ranked query (EIRQ) schemes based on the aggregation and distribution layer (ADL) introduced. It is used for aggregating user queries and distributing search results. Under the ADL, the computation cost incurred on the cloud can be largely reduced, since the cloud only needs to execute a combined query once, no matter how many users are executing queries. And also provide differential query services while protecting user privacy. By using our schemes, a user can retrieve different percentages of matched files by specifying queries of different ranks. By further reducing the communication cost incurred on the cloud, the EIRQ schemes make the private searching technique more applicable to a cost-efficient cloud environment. However, in the EIRQ schemes, we simply determine the rank of each file by the highest rank of queries it matches.

ACKNOWLEDGMENT

It facilitates partners to make available services that use the Cloud infrastructure. The Solution EES program is intended to help Partners augment their cloud contribution and widen experience to new customers and global markets. EES offers software, applications and cloud services on top of public Cloud. The program consists of a set of tools such as platform as a service (PaaS) and software as a service (SaaS) vendors provide cloud based services include Application, Database, Development & Testing, Management e.g. Orchestration, Mobile computing, Monitoring, Multimedia, Platform as a Service, Security, Storage and Technology.

REFERENCES

1. P. Mell And T. Grance, ‘‘The Nist Definition Of Cloud Computing (Draft),’’ In Nist Special Publication. Gaithersburg, Md, Usa: National Institute Of Standards And Technology, 2011.
2. R. Curtmola, J. Garay, S. Kamara, And R. Ostrovsky, ‘‘Searchable Symmetric Encryption: Improved Definitions And Efficient Constructions,’’ In Proc. Acm Ccs, 2006, Pp. 79-88.
3. J. Bethencourt, D. Song, And B. Waters, ‘‘New Constructions And Practical Applications For Private Stream Searching,’’ In Proc. Ieee Sp, 2006, Pp. 1-6.
4. Q. Liu, C. Tan, J. Wu, And G. Wang, ‘‘Cooperative Private Searching In Clouds,’’ J. Parallel Distrib. Comput., Vol. 72, No. 8,Pp. 1019-1031, Aug. 2012.
5. “Space-efficient private search with applications to rateless codes”, Financial Cryptography and Data Security, 2007. [10] M. Finiasz, K. Ramchandran,“Private stream search at the same communication cost as a regular search: Role of ldpc codes”, In Proc. of IEEE ISIT, 2012.
6. X. Yi, E. Bertino,“Private searching for single and conjunctive keywords on streaming data”, In Proc. of ACM Workshop on Privacy in the Electronic Society, 2011.
7. B. Hore, E.-C. Chang, M. H. Diallo, S. Mehrotra, “Indexing encrypted documents for supporting efficient keyword search”, in Secure Data Management, 2012.
8. P. Paillier,“Public-key cryptosystems based on composite degree residuosity classes”, In Proc. of EUROCRYPT, 1999.

