Secure Multi-keyword Search based on TF Technique

Amol D.Bhand[1], Prof. Sachin.K.Korde[2]
[1, 2] Department of Computer Engineering
[1, 2] PREC, Loni, Maharashtra.

Abstract- *Cloud computing has more important new model of enterprise IT infrastructure, in which it can organize large resource of computing, storage and various applications, and enable users to enjoy everywhere, convenient and on-demand network access to a shared pool of configurable computing resources with great efficiency and minimal economic requirements. These interesting features, personal users as well as industries are forced to put their data to the cloud, instead of buying hardware and software to handle the data yourself.*

In this paper, Secure Multi-Keyword Search Based on TF Technique, which can supports operations like update delete and insertion of documents as the authorized user want. A tree-based index structure is developed and Greedy Depth-first Search algorithm to give efficient multi keyword ranked search is added in this scheme. The kNN algorithm which is used to encrypt the query vectors and index along with that can gives accurate relevance score calculation amoung encrypted index and query vectors.TF-IDF algorithm is help to give the results of uploading the multiple files over cloud and how much repeated words will occurs which will helpful for ranking the collection of documents. The use of the special tree-based index schema, the proposed framework can get sub-linear search time and deal with the deletion and insertion of documents flexibly.TF-IDF helps get to the user for giving accurate result.

Keywords- Encryption, Multi-keyword ranked search, Dynamic update, Cloud computing.

I. INTRODUCTION

Cloud computing, is the most important part in the computer world that can gives a various data for processing along with other devices also taking part of that. It is a model for enabling all resources, any time any where access to a shared pool of configurable computing sources (e.g. applications services, servers, networks and storage).

Despite of the different advantages of cloud services, outsourcing important information to remote servers brings privacy area. The cloud service providers (CSPs) that keep the data for users may access users' sensitive information without authorization. A commonly use method to secure the data security is to encrypt the data before moved to cloud server. While ,this will cause a large cost in terms of data usefulness. In the previous techniques keyword-based information retrieval, which are broadly used on the plaintext data, cannot be directly applied on the encrypted data. Firstly download all the data from the cloud server which we have want and decrypt it on your local machine.

To overcome the above problem by introducing the searchable encryption (SE) schemas have made specific contributions in terms of efficiency, functionality and security. SE scheme is used to give the client to store up the encrypted data (unreadable data) to the cloud and implement keyword search over ciphertext domain. To achieve various search functionality some thread model proposed, such as single keyword search, similarity search, multi-keyword Boolean search, ranked search, multi-keyword ranked search. Out of that, multi-keyword ranked search received more and more attention for its practical applications. Now a days, some dynamic methods have been developed to support inserting and deleting operations on document collection.

In this paper, a secure tree-based search scheme over the encrypted cloud data, that supports multi-keyword ranked search and operation on the document collection is developed as well as implemented.

The "term frequency(TF)×inverse document frequency (IDF)" model and the vector space model are jointly to used in the index construction and query generation to provide multi-keyword ranked search. In order to obtain highest search efficiency, we developed a tree-based index structure and propose a "Greedy Depth first Search(GDFS)" algorithm on the basis of index tree. The use of structure of tree-based index, the search system can flexible to gives sub-linear search time, deletion and insertion of documents. The secure kNN algorithm is helps to encrypt the index and query vector, and as the time to give accurate relevance score calculation between encrypted index and query vectors. To avoid different threat attacks in different models, we create secure search scheme: the basic dynamic multi keyword ranked search (BDMRS) scheme in the known ciphertext model, and which will executed on background model, enhanced dynamic multi-keyword ranked search(EDMRS)

known background model which will be the role in this system.

II. BACKGROUND

A. Cloud Computing

The cloud computing technology deals with configuring, accessing and manipulating the software and hardware resources remotely. It is beneficial to the end user by giving online data storage, infrastructure, and application. Since management of data and infrastructure in cloud is handled by third-party, it is danger to provide the sensitive information to cloud service providers(CSPs). There is a number of technologies making cloud computing flexible, reliable, and usable. Some of them include virtualization, service-oriented architecture, grid computation etc. Encryption will help to protect data from being compromised. Although encryption provides security to the data from any unlawful access, it does not prevent data loss.

B. Searchable Encryption Scheme

The searchable encryption scheme falls to two types: Symmetric searchable encryption and asymmetric Searchable encryption. In private-key searchable encryption, the user encrypts the data with the private key and can use some data structures to make accessing relevant data efficiently. The data and the data structures are encrypted and store on the server so that only user will have the private key only they can access it. Here the starting work for pre-processing the data is very large, but later work for accessing the data is very small. Public-key searchable encryption(SE) will involve both public key and private key. The owner encrypts the data with public key and outsources it to the server. Public key as well as private key was generated by the owner. The users having public key can add words to index but only the user who knows the private key can generate trapdoors and performs search.

III. RELATED WORK

The SE schemes facilitate the clients to store the encrypted data to the cloud and execute secure keyword search over cipher text domain. The single keyword SE problem is analyzed by Song et al. [1] for electronic message systems in symmetric setting. This scheme has no index, therefore the search operation performed for the complete file and no ranking operation upon the result document. Secure index made by Bloom filter in [9] is proposed by Goh et al. Curtmola et al. gave the formal definition of the searchable encryption and developed an index scheme depends on the inverted list in [8]. In , Wang et al. solved the problem of ranking with the keyword frequency and order-preserving encryption. The first SE scheme created by Boneh et al. [3] uses asymmetric encryption. In this, the searching time is linear to the collection data and no ranking methods executed. Naveed et al. [4] construct a storage system which achieves searchable encryption by access pattern of search user. Data files are divided into blocks and blocks are indexed. All of these works are based on the single keyword search over the encrypted data.

The search is modified by creating scheme that supporting conjunctive keyword search [5-6]. The proposed privacy-preserving multi-keyword ranked search scheme by Cao et al. [4], has huge measurement of overhead as it needed calculation of relevance score for all document. Sun et al. [2] proposed privacy-preserving multi-keyword rank search cosine similarity method.

Li et al. proposed a fuzzy search over encrypted data in [10]. Then the improved the scheme by reducing the size of index. In improved [10] by introducing a tree structure index and enriched the search functionality.

IV. EXISTING SYSTEM

A basic approach to secure the data security is to encrypt the data before uploading over cloud. SE schemes enable the data user to store the encrypted data to the cloud and execute keyword search over ciphertext domain. So lots of works have been proposed under different threat models to achieve various search functionality, such as single keyword search, similarity search, multi-keyword boolean search, ranked search, multi-keyword ranked search, etc. Among them, multi-keyword ranked search gets more and more interest for its useful applications. Now a days, some dynamic schemes have been developed to support insertion and deleting operations on document collection. These are important works as it is highly possible that the data owners need to update their data on the cloud server.

Disadvantages

1. Large expenditure in terms of data usability. For example, the existing system on keyword-based information retrieval, which are widely used on the plaintext data, cannot be directly applied on the encrypted data. Downloading all the data from the cloud and decrypt it on local system is impractical.
2. Existing System method not practical due to their high computational overhead for both the cloud sever and user.

V. PROBLEM STATEMENT

Design a system for secularly search over cloud in the form of encrypted data. System will be

1. Rank based search data.
2. Higher efficiency will be given.
3. Top k rank result.
4. On the basis of TF IDF algorithm.
5. Resist from different attack.
6. How to Store data in Securely over cloud.

VI. SYSTEM COMPONENETS

The implementation of this work is carried out in four modules. Which will be

A. Data Owner Module.
B. Data User Module.
C. Cloud Server Encryption Module.
D. Trapdoor.

Secure Multi-keyword Search based on TF Technique scheme has the following.

Data Owner

The Data owner has a collection of documents D={d1,d2,…...,dn} which to be forwarded in encrypted form to the cloud server. The data owner creates a secure searchable tree index I from document collection D, using some preprocessing IR operations and then creates an encrypted document collection. Then, the data owner forward both the encrypted collection C and the secure index I to the cloud server and allow secure search by the user. While data user is want to access the data of data owner it will take the permission of data owner by using sharing keys or some trusted methodology. Data owner has a authority to revoke the any user and handling the various files between this group it will depends on owner.

Data User

Data users are authorized ones to access the documents of data owner between the created groups of data owner. While data user is want to access the data it will be authorized by data owner then and then only data user will access the data of data owner added with particular of that group. While performing the search, the authorized user can generate keyword which he/she want the particular keyword data user will search it and sent it over the cloud server. The user can view the top k encrypted documents and if the user satisfying access policy and wish to download, can decrypt the documents with the shared secret key from data owner. Then and then only data user can access the data of data owner.

Cloud server

Cloud server stores the encrypted document collection C and the secure searchable index tree I for the data owner. Upon receiving the request from the data user, the cloud server performs search over the index tree I, and returns the top-k relevant encrypted documents to the data user.

Trapdoor

Trapdoor can stores the TF-IDF value which can be generated with the help uploading the number of documents over the cloud by data owner. By TF-IDF algorithm it can show the TF,IDF values of particular documents. This operations mainly done over the cloud it will be depends on uploading data owner documents that's why the values will be change accordingly. By using standard TF-IDF algorithm it can be shown the particular results.

This system mainly done the operations on .TXT files only, But you can upload any type file (pdf, docx,any type images, etc) with any size and all the data stored on cloud is an encrypted format.

VII. SYSTEM ARCHITECTURE

Architecture of system is diagrammatically shown below in diagram (fig. 1)

In the below diagram we will try show how the system will work and what will be the flow of particular system.

Architecture of Secure Multi-Keyword Search Based on TF Technique consist of four module which will be Data owner, Data user, cloud server, trapdoor. But trapdoor is basically background model and not broadly explains the work. The system will also have a third party auditor to provide authentication and verification. The user is logged in using attributes that is given by data owner.

The system consists of following four operations i.e., File upload, Authentication, search and ranking.

A. File upload

While uploading the file data owner site creates the file and its index in tree based. First he/she selects the file and performs some IR processing such as tokenization etc. Also calculate the TF and IDF values of each data item which will be shown in trapdoor module. After measuring the frequency of tokens, they are sorted in descending and most frequent tokens are chosen as keywords for search. These keywords are added to the index table in sorted order. In encryption module, setup() output the private key SK and public key PK. The file F and index I is encrypted at data owner side and uploaded to the server. The encrypted keywords are inserted into the b-tree for efficient search.

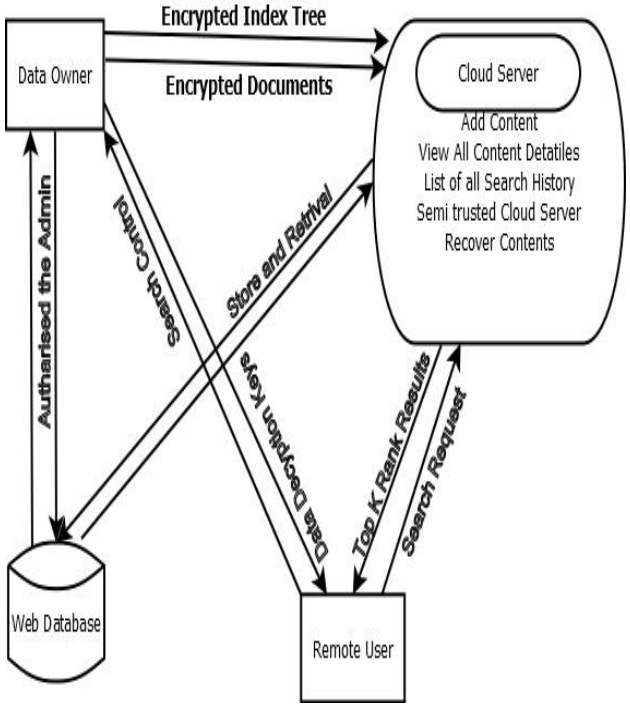

Figure 1. Architecture Diagram

B. Authentication

Authentication procedure is run by the admin/data owner. The user sends the authentication parameters (such as a user name and password) to the data owner. The data owner verifies the details and identify whether the user is authorized or not. The user is register by giving various credentials l like name, password, email etc. and these attributes are used define access policies for the user and data owner will decide that the user will added in which group. Before uploading, owner verify the content of the particular file. After executing the request, the admin/data owner will upload the file in which the data owner is create the group of registered user and only for that particular group members can use the file. This ensures that only authorized user by data owner can access the file. Data owner and user share a secret key. Then user can decrypt the file by using private key .

C. Search

The data user has functions such as user registration, keyword search and file download, uploading its own data for his/her personal use. The user can register by giving his/her credentials such as name, password, mail id, phone number etc. In keyword searching function, the user can specify multiple keywords, also the number of relevant documents he/she need or that contains the query word will appears. User creates request by encrypting the multiple keywords for which search to be performed. The data user generates secure Q is the user query and PK is the Primary key which user send it to the server.

D. Ranking

Document ranking is done by calculating the cosine similarity between the query vector and document. The cosine measure uses TF×IDF rule, where TF denotes the occurrence count of a term within a document (high TF means the term highly correlated to a particular document), and IDF is obtained by dividing the total number of documents in the collection by the number of documents containing the term.

In this paper Secure Multi-Keyword Search Based on TF Technique, which supports multi-keyword ranked search and dynamic operation on the document collection. The widely-used "term frequency (TF) × inverse document frequency (IDF)" model and vector space model are jointly in the query generation and index construction to survives multi-keyword ranked search. In order to get exact or large search efficiency, we construct a tree-based index structure and propose a "Greedy Depth-first Search" algorithm based on this index tree. The secure kNN algorithm is used to encrypt the index and query vectors, and for the time being ensure accurate relevance score calculation between encrypted index and query vectors. To resist different attacks in different threat model, we construct secure search scheme: the basic dynamic multi-keyword ranked search (BDMRS) scheme in the known ciphertext model, and enhanced dynamic multi-keyword ranked search(EDMRS) known background model which will be the role in this project.

VIII. FLOW OF SYSTEM

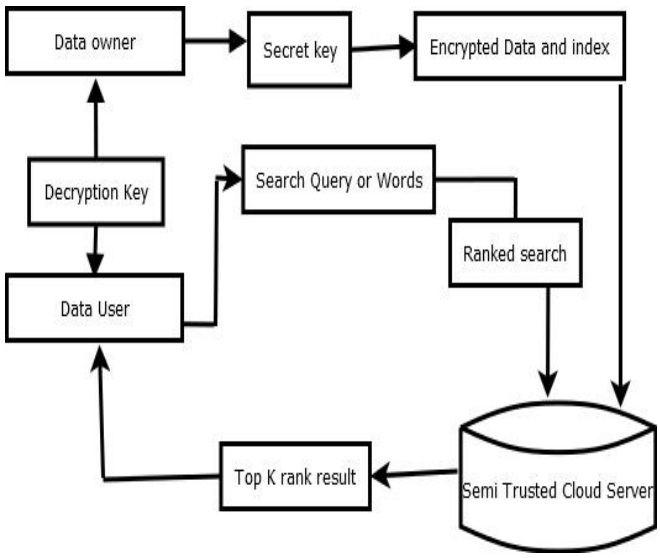

Figure 2. flow of system

In the above figure that can be show the flow of the system. For the implementing system that can be used the various algorithms in that, for the encryption and decryption can used the AES algorithm will be used which will be size of block 128 bits and key length will be 128, 192, 256 bits. Also BDMRS scheme and kNN algorithms takes important part for ranking the documents giving to user.

TF-IDF will gives that calculating the relevance score and top k ranked search will have to give appropriately.

IX. RESULTS AND DISCUSSION

In Secure Multi-Keyword Search Based on TF Technique is the system will be constructed and giving the various results because of using various algorithms like greedy depth first search(GDFS),for constructing index we will use the tree algorithm, KNN algorithm, TF-IDF algorithm will take the important part in our schema and some other algorithms. and for the security purpose we will use the encryption and decryption and for that algorithms like HMAC, RSA, etc. This algorithms is will be the part of cryptography area and we will use for the security purpose

Therefore by using this various algorithms and then observing the system we will get the some results and which will be explained below

For the another results and improvement of this project will be under construction at this stage we will shown this much results only.

A. Efficiency of Index tree construction

The process of index tree construction for document collection F include two step 1.Building the unencrypted index tree of the document collection of files F proceeding to encrypted index tree by tokenization and multiplication of matrices (m*m).

The time cost for creating the index tree is depends on the how much documents are upload the data owner over cloud in document collection F and number of keyword will added in the dictionary for the use searching and ranking of the particular document. Fig 2 shows that the how much time cost will take for the index tree construction

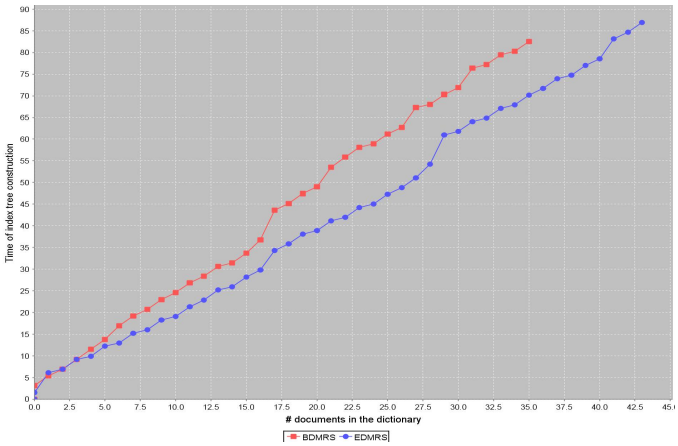

Figure 3. Time cost for index tree construction

In the fig.3 that can show that ,the user which want to Use the data of data owner .The authorized user is search on it. The system will give the results on the basis of its TF count.

In the fig 4 can show that trapdoor part can show that the TF of particular file that can how many various keyword are there in that file which will be calculated and it will display on descending order of particular words.

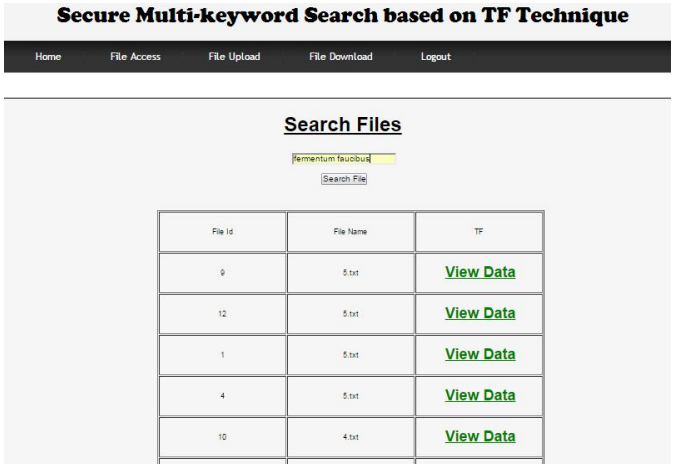

Figure 4. Searched file on the basis of its TF value

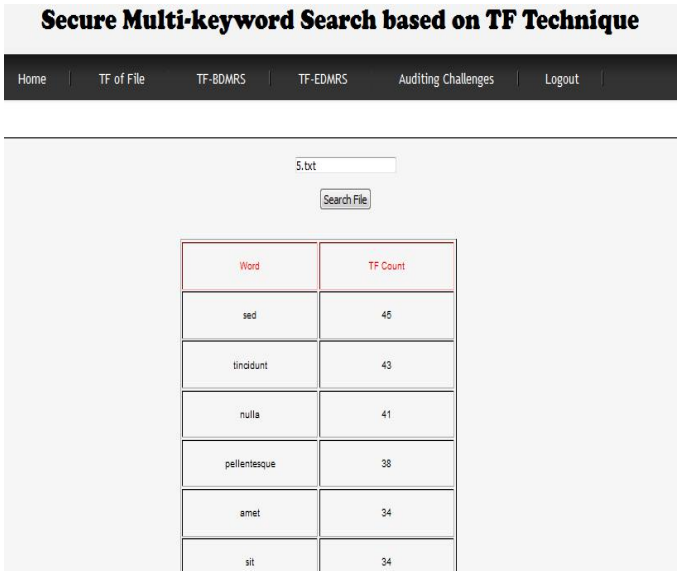

Figure 5. Count of word in particular file

In fig 5 can show that the calculation of BDMRS and EDMRS scheme

In the below fig 5 show the TF-IDF on basis of BDMRS scheme and showing the results by using this schemes only

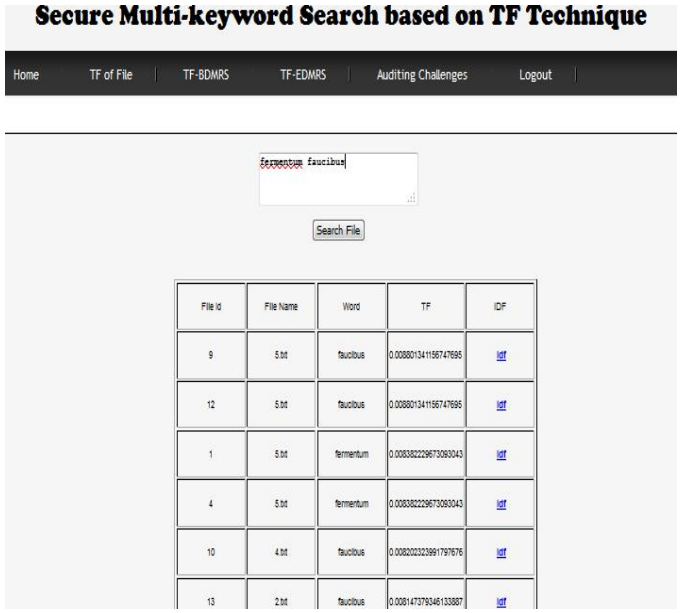

Figure 6. TF-IDF calculation on the basis of BDMRS Scheme

X. CONCLUSION

In this paper, we implement a special keyword balanced binary tree as the index, and propose a Greedy Depth-first Search algorithm to obtain better efficiency than linear search. The parallel search procedure can be done to additional reduce the time cost. The security of this proposal is protected against two models by using the secure kNN algorithm. In this project, the data owner is dependable for creating, updating information and sending them to the cloud server. TF-IDF model will help that finding values which will helps for trapdoor generation Thus, the data owner needs to store the unencrypted index tree.

XI. ACKNOWLEDGMENT

I would like to thank my project guide Prof. S. K. Korde, for her personal involvement and constructive suggestion throughout the work. I would like to thank to PG Coordinator Prof. S.D.Jondhale, HOD Prof. S.M.Rokade and our Principal Dr. Y.R.Kharde for providing me an opportunity to work on this topic and his valuable support for my work.

REFERENCES

[1] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. IEEE Symp. Secur. Privacy,2000, pp. 44–55.

[2] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li,"Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking," in Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. secur., 2013, pp. 71–82.

[3] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in Proc. Adv. Cryptol.-Eurocrypt,2004, pp. 506–522.

[4] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," in Proc. IEEE INFOCOM, Apr. 2011, pp. 829–837.

[5] Y. H. Hwang and P. J. Lee, "Public key encryption with conjunctive keyword search and its extension to a multi-user system," in Proc. 1st Int. Conf. Pairing-Based Cryptography, 2007, pp. 2–22.

[6] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in Proc. 4th Conf. Theory Cryptography, 2007,pp. 535–554.

[7] P. Golle, J. Staddon, and B. Waters, "Secure conjunctive keyword search over encrypted data," in Proc. Appl. Cryptography Netw. Secur., 2004, pp. 31–45.

[8] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: Improved definitions and efficient constructions,"in Proc. 13th ACM Conf. Comput. Commun. Secur.,2006, pp. 79–88.

[9] E.-J. Goh, "Secure indexes," IACR Cryptol. ePrint Archive, vol. 2003,p. 216, 2003.

[10]J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in IEEEProc. INFOCOM, 2010, pp. 1–5.

[11]Zhihua Xia, Xinhui Wang, Xingming Sun,Qian Wang, A Secure And Dynamic Multi-Keyword Ranked Search Scheme Over Encrypted Cloud Data IEEE Transactions On Parallel And Distributed Systems, Vol. 27, No. 2, February 2016.