A Novel Approach in Cloud for Secure Authorized De-Duplication

Putta Kalyan Kumar[1], B.Venkaiah Chowdary[2]

Putta Kalyan Kumar pursuing M.Tech (CSE), Chintalapudi Engineering College, Ponnur. A.P., India.

Bhimineni Venkaiah Chowdary, working as Asst. Professor, Department of Computer Science & Engineering, Chintalapudi Engineering College, Ponnur. A.P., India.

Abstract: *Data deduplication is the mechanism using which the data storage process can be improved. It is one of the compressions like technique where by using this mechanism we can limit the storage and efficiently utilize the database storage capacity. Deduplication is the process of eliminating the similar kind of data from the storage device to reduce the amount of storage space and save bandwidth. For the security reasons, to save data stored at cloud we need to encrypt the data and provide security key with which the data can be used by the end user. Deduplication can be achieved in three ways file level, content level and byte level. In this work we would like to work on file level deduplication, content level deduplication and around 50% of byte level deduplication. The complete mechanism of deduplication will be carried out in the cloud architecture.*

Keywords: *Deduplication, Compressions, Encryption, Cloud Architecture.*

INTRODUCTION

Cloud Computing is the technology which makes data available to the end user at any given point of time and can store huge amount of data. Size of the cloud cannot be decided it is very vast. As the size is big, in the same way the cost incurred to maintain the data in the cloud is also considered to be high because of the service made available by the cloud providers. There is a term called TPA (Third Party Auditor) the role of TPA is to authenticate the user logging into the application and provide security to the cloud from intruders. Intruder is a user who tries to access the data from the system without authenticating self or trying to do some crime by taking the data from the cloud. Cloud providers provide the space to the companies to upload their data into cloud and make it available to their clients whenever it is needed for them. The main reason for going to cloud architecture is that whenever the company has got more clients and the data to be served to the clients is more which cannot be withhold by server then comes into picture the external storage which refers to Cloud. Cloud is the database storage which is provided by some other organizations and as many companies depend on Cloud there is an authenticator called TPA who will validate the entries into Cloud and will not allow the users who are non-genuine.

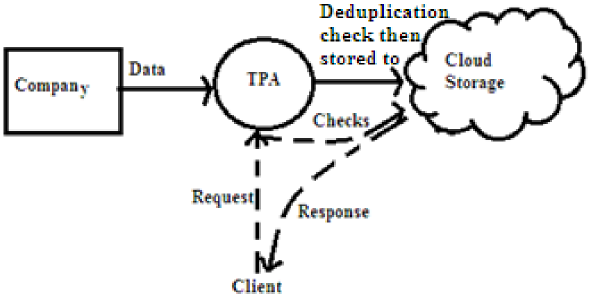

Fig 1: Cloud work flow

As shown above we can see that the company and Client both are connected to Cloud via TPA. The role of TPA is very important in the Cloud Computing processing. The point that arises is whether the data uploaded into cloud is safe or not, for this purpose the setup is coming up with the implementation of encryption algorithms for the data that is made available in the cloud just to make sure that data reaches the correct person but not the intruders.

To tell about the real instances of Cloud computing we use many applications in day to day usage, few of which are discussed below. Consider the Gmail application which is very common these days for mailing purpose and in this application we will go with communication between two people i.e. a person has send a mail with excel sheet attached and the user who received that file does not have any MS-Office in the system to download and read it, but without even downloading it the user is given the flexibility to read the data without downloading and this is possible because of Cloud Computing which is employed with the Gmail application.

Second very famous application which almost many of the users use is Facebook, a social network. This application is used by many people across globe and so the database required for this is very huge and a single server cannot withstand this many users and their data. The other thing that can be noticed here is that a user of facebook can have any id for the login i.e. Yahoo id or Gmail id or Rediff id is accepted as a username, the point to be understood is that all the servers of various companies are kept in a single room like architecture called Cloud and because of which a genuine user of any of these domains can easily access even Facebook application.

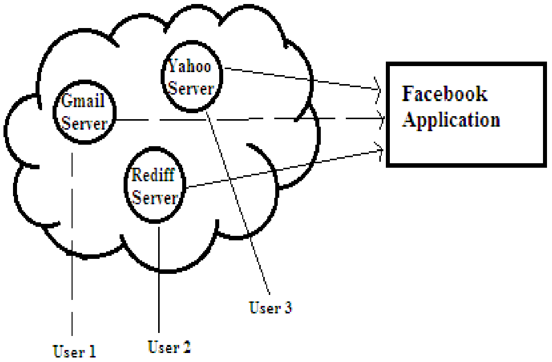

Fig 2: Social community Facebook Architecture

BACKGROUND

Here we are going to discuss about the background work being done in the cloud architecture to have a secure authenticated deduplication process. For security we are using plain cipher encryption because our main concern is not the security, we are majorly focusing on the deduplication architecture to efficiently utilize the cloud resources. Deduplication is generally a protocol that when followed by any organization will eliminate the duplicates in the data. Deduplication is implemented in three ways i.e. file level, content level and byte level. Coming with the file level deduplication, it will check the file names whether the file was already uploaded by the user or not. Content level deduplication is a type of approach using which we can check for duplicate data in the complete content i.e. it will check the complete data by considering it as string. Third type of check is byte level deduplication i.e. it will check the content byte by byte and it becomes clumsier while performing this type of deduplication. In our approach, we are implementing the file level, content level and 50% check under byte level deduplication. This concept of comparing the data in the file is done taking the help of io package in java language. There are classes like FileInputStream, FileReader which are used to read the data from the file and coming with the utility package of java language, it is giving support to store the data temporarily and perform the deduplication check. Classes that are used under the utility package are HashSet, TreeSet and HashMap.

The above mentioned classes play a vital role in the core functionality and coming with the data that need to be stored in the server should be provided security as it is the area which is seen by almost all the companies which store their data in the same area. To overcome the drawback of the losing the data in server or being traced by intruder we are using the plain cipher encryption to safeguard the data. For securing the data we firstly need to generate key for the data that we are going to encrypt and to do so we are taking the help of utility package of java. In this package we have a class with the name ‘Random’. This class is giving us the flexibility to generate key and the key size here can be decided by the owner. For our approach we are using the key size of 6bits. Below is the sample code for generating the key of 6 bits.

```
String
ss="abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMN
OPQRSTUVWXYZ0123456789";

String key="";
        int i=0;

java.util.Random r=new java.util.Random();

        for(i=0;i<6;i++){

key=key+ss.charAt(r.nextInt(ss.length()));

        }
```

this above process a key will be generated and it can be used for when downloading the data from the server and it will be shared among all company users.

After generating the key, we need to encrypt the data which is converting the plain text to cipher or unreadable format; just to make sure that the data even tough taken by any user cannot be utilized for any further purpose. Below is the sample code to show the process of both encryption and decryption using plain cipher cryptography technique,

```
public static String encrypt(String sa)
{
int ch;
for(int i=0;i<sa.length();i++)
{
ch=sa.charAt(i)*16;
finaldata=finaldata+ch+'#';
}
System.out.println("EncryptedData:"+finaldata);
}
return finaldata;
}

public static String decrypt(String finaldata)
{
```

```
String dat=finaldata;
char c;
StringTokenizerdatt=new StringTokenizer(dat,"#");
while(datt.hasMoreTokens())
{
String s=datt.nextToken();
int i=Integer.parseInt(s);
i=i/16;
c=(char)i;
finaldat=finaldat+c;
}
System.out.println("Decrypted file is:"+finaldat);
return finaldat;
}
```

In the above snippet we can see the process of both the encryption and decryption. During the process of encryption, the complete data is taken into string and this string is broken and each character that is coming out is converted to its ascii initially and then that value is multiplied by a integer ‘16’, by doing so a new value is generated which when taken by any user cannot be converted back to actual value anytime. During the decryption phase, the value which was generated in the encryption phase is broken down based on the special character that was added to the final string and that value is then divided by the integer ‘16’ only when the user gives the actual key. In this manner this process is repeated till the end of the string is reached. This completes the decryption phase and we get back the final string. In this manner we are achieving the security for the data.

Coming with the deduplication process, it is the main theme in our proposed them, as already said that the deduplication theme in our work is purely supported by the io package of java. Algorithm for the deduplication check will be in the following manner,

Steps for De-duplication check:

1. Check database if no file, store the file into database table.
2. If file exists, then check for file name
    a. If file name exists, then check the content taking into a string format with the already existing data.
    b. If file name exists but with different content then new row will be inserted by modifying the file name by adding some extra characters.
3. If file name does not exist, then it considers that the data is new and file directly gets added to database.
4. In above steps, it is showing the de-duplication for both filename level and content level.
5. If the file name is same and then the content will be verified byte by byte and we are considering only 50% byte level de-duplication to make us understand the actual process of deduplication and to know the importance of it in the real time servers.

CONCLUSIOSN AND FUTURE WORK

Paper addressesto the problem of deduplication and thereby with the implementation of the solution to this problem we are overcoming with the issue of misuse of the server resources that actually need to be utilized in an efficient manner. Our work shows the deduplication implementation process in three ways; file level deduplication, content level deduplication and byte level deduplication. It is very important concept in the real time servers as the space allocated to any domain must be utilized in an efficient manner and thus should not misuse the space allotted in the server or cloud.

Our future scope of work would be on the security issue in addition to the deduplication process. We would propose AES cryptographic algorithm for securing information to ensure the data stored in the server cannot be misused by any unknown users accessing the server/cloud data.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, “Mining AssociationRules between Sets of Items in Large Databases,” ACM SIGMODRecord, vol. 22, pp. 207-216, 1993.

[2] R. Agrawal and G. Psaila, “Active Data Mining,” Proc.First Int’lConf. Knowledge Discovery and Data Mining, pp. 3-8, 1995.

[3] R. Agrawal and R. Srikant, “Mining Generalized AssociationRules,” Proc. 21th Int’l Conf. Very Large Data Bases (VLDB ’95),pp. 407-419, 1995.

[4] M.L. Antonie, O.R. Zaiane, and A. Coman, “Application of DataMining Techniques for Medical Image Classification,” Proc.SecondInt’l Workshop Multimedia Data Mining (MDM/KDD ’01), 2001.

[5] W.-H. Au and K.C.C. Chan, “Mining Changes in AssociationRules: A Fuzzy Approach,” Fuzzy Sets Systems, vol. 149, pp. 87-104, Jan. 2005.

[6] E. Baralis, L. Cagliero, T. Cerquitelli, V. D’Elia, and P. Garza,“Support Driven Opportunistic Aggregation for GeneralizedItemsetExtraction,” Proc. IEEE Fifth Int’l Conf. IntelligentSystems (IS ’10), 2010.

[7] S. Baron, M. Spiliopoulou, and O. Gnther, “Efficient Monitoring ofPatterns in Data Mining Environments,” Advances in Databases andInformation Systems, L. Kalinichenko, R. Manthey, B. Thalheim,and U. Wloka, eds., vol. 2798, pp. 253

Authors Profile

PUTTA.KALYAN KUMAR,

Pursuing M.Tech In Chintalapudi Engineering College, Ponnur, Department Of Computer Science & EngineeringMy research interests are network security.

BHIMINENI.VENKAIAH CHOWDARY, M.TECH Assistant Professor Chintalapudi Engineering College, Ponnur Deparment Of Computer Science & Engineering, and has 4 Years of Teaching Experience.