An Approach to Improve Cloud Data Privacy by Preventing from Data Mining Based Attacks

K.Sasireka[1]
Associate Professor, Dept of Information Technology
Narasu's Sarathy Institute of Technology
Salem, India
keerthisasi2007@gmail.com

Dr.K.Raja[2]
Dean (Academics)
Alpha College of Engineering
Chennai, India
raja_koth@yahoo.co.in

Abstract- Cloud Computing provides a good model for the providers to deploy the computing infrastructure and applications on-demand. It offers greater flexibility to users by connecting to various computing resources and allowing access to IT enabled services. But it has the risk of privacy of user data and security. Thus security among the users of cloud is the most important concern. One of the security issue in cloud computing is data mining based attacks, which involves that the data can be analyzed continuously by an unanonymous person to get the valuable information. Using the single cloud provider this is a major problem among the clients in the cloud, because the outside attacker can analyze their data for a long time to gain the sensitive information. In this paper, we have given the data mining based attacks on cloud data and a method to prevent the attacks.

Index Terms- Cloud Computing, Data Mining, Security, Privacy etc.

INTRODUCTION

Cloud computing enables the end-users, small and medium-sized companies to access computational resources like storage, software etc. In cloud computing, with these vast amount of computing resources, users are able to solve their problems easily using the resources provided by cloud. Some of the cloud services include Software as a Service(SaaS), Platform as a Service(PaaS), Infrastructure as a Service(IaaS)[2]. The examples of cloud services provided by big organizations are: Elastic Compute Cloud(EC2) by Amazon, Google App Engine(GAE) by Google and SQL Azure by Microsoft etc.

Using the cloud computing services, users are able to store their data in servers and access their data from anywhere and they need not worry about the lose of data due to disk faults, system breakdown etc. But there are several security issues in cloud like assurance and confidentiality of user data. The users who are entrusting the cloud provider may lose the access to his data either permanently or temporarily due to any unexpected event like malware attack. This unexpected event provides significant harm to the users. The providers in cloud can analyze the user data continuously and similarly the outside attackers who try to get access to the cloud can also analyze the user data. So, the user may lose his data privacy.

There are various data analysis techniques are available now to extract the sensitive information from cloud data. The outside attackers can use these techniques to get the sensitive information from cloud[10]. The potential threat to cloud security may be data mining where the large volume of data belonging to a particular user will be stored in a single cloud provider. This single cloud provider approach is the main drawback in cloud where the provider can use more powerful data mining algorithms to extract the private information of user. The second drawback of this approach is the attackers who have unauthorized access to the cloud can use the data mining techniques to extract the sensitive information in the user data.

In this paper, we present a approach to provide unique identity to the cloud users and servers known as Federated Identity Management and to prevent data mining attacks by using multiple cloud providers. The user data will be distributed among multiple cloud providers, so it will be a difficult task to the attackers to get the data. The key idea of our approach is to classify the user data, divide the data into small chunks and distribute these chunks to the various cloud providers. Simply, this approach consists of 3 steps: classification, fragmentation and distribution of data. Classification is a process where sensitive data is identified and appropriate mechanisms are implemented to maintain privacy of this sensitive data. Fragmentation is a process where the data is divided into small chunks. Distribution is a process where the divided chunks will be distributed to cloud providers. Distribution of data to a cloud provider can be done depending upon the reliability of cloud provider and data sensitivity. The reliability of a cloud provider means if the cloud provider is able to store the data chunks with such sensitivity. Using this approach, it is difficult for the attacker to get the data chunks from different providers and also mining sensitive information from these data chunks is a tedious process[8][9].

RELATIONSHIP BETWEEN CLOUD COMPUTING AND DATA MINING

Data Mining is the major growing field in IT industry which is also known as Knowledge Discovery in Databases(KDD)[1]. It is used to discover useful patterns from large volumes of data. In data mining, the main areas are like Frequent Pattern Mining, Association Rule Mining etc.

Cloud Computing and Data Mining are closely related to each other. The interrelationship between these two is having its advantages and disadvantages. The advantage is: data mining has been used by cloud providers to provide better service to clients. The disadvantage is: attackers outside the cloud provider who is not having authorized access to cloud, will also use data mining to extract data from cloud. The extraction of useful data from cloud involves 2 factors: suitable amount of data and appropriate mining algorithms. There are so many mining algorithms which will work good to extract useful information from cloud which violated the user data privacy. For example, association rule mining algorithms[3] can be used to find association relationships between huge number of business transaction records etc. Thus data mining is

becoming a powerful tool and possess more threats to cloud users.

DATA MINING ATTACKS IN CLOUD

The main vulnerability in the present cloud storage system[7] is data reside in a single cloud provider which leads to data loss due to malware attack, network disruption, cloud provider going out of business etc. This is illustrated in Figure 1.The attacker may be an inside attacker(malicious persons at cloud provider) or an outside attacker(persons outside cloud provider). If an attacker wants to attack a particular client, he may aim at the single cloud provider and gather the sensitive information of that client data. Thus this single provider system in cloud is a greatest security concern. Both the inside and outside attackers have the advantage of using data mining to extract the useful data.

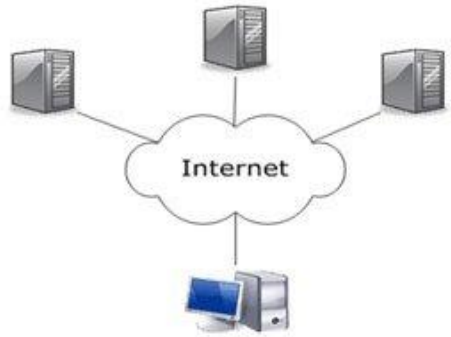

Figure 1. A Simple Cloud Architecture

Our proposed system consists of multiple providers and data can be splitted into several chunks and distributed among the providers[6]. Thus distributing the data to multiple providers prevents the cloud provider from accessing all the chunks of a client. Even if he performs mining on the chunks to the provider, the extracted knowledge will be incomplete which leads to unsuccessful mining[5].

This method can be implemented in traditional databases using Redundant Array of Independent Disks (RAID) technique. Redundant Arrays of Cloud Storage (RACS) uses this technique to reduce the cost of maintaining the data in cloud[4]. Different RAID levels can be choosen to ensure the assurance of data. Higher RAID level leads to higher level of assurance. The main advantages of this approach are:first privacy is improved by increasing the number of destinations and second amount of data stored at each destination is reduced.

SYSTEM ARCHITECTURE

Our proposed system is illustrated using Figure 2 as shown below.

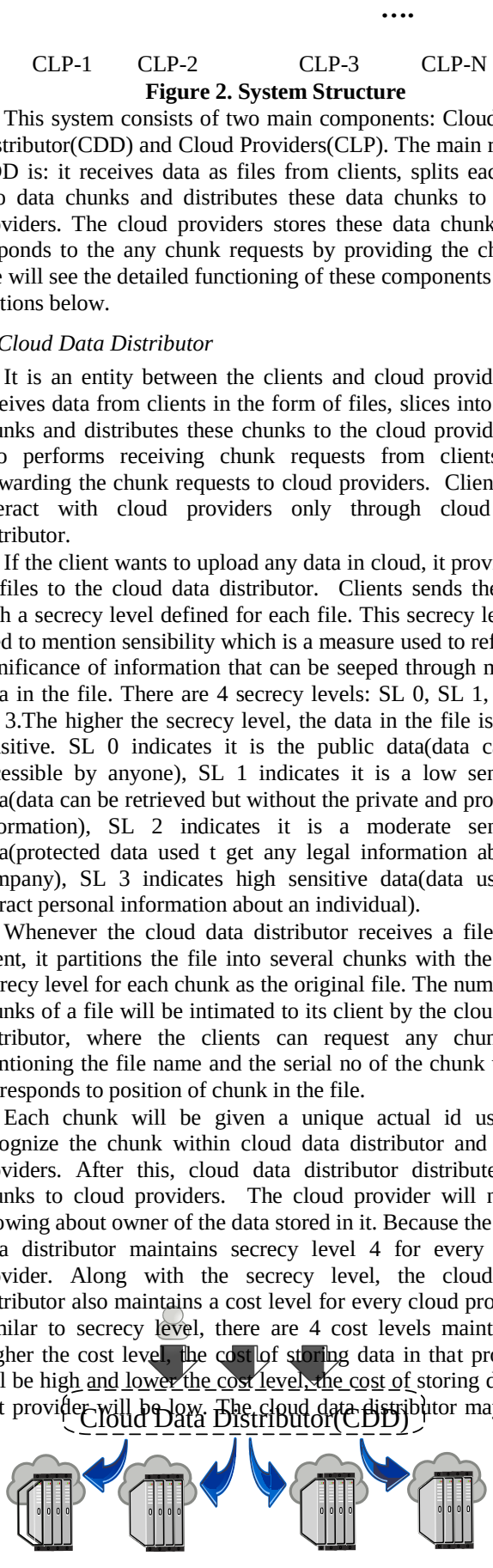

Figure 2. System Structure

This system consists of two main components: Cloud Data Distributor(CDD) and Cloud Providers(CLP). The main role of CDD is: it receives data as files from clients, splits each file into data chunks and distributes these data chunks to cloud providers. The cloud providers stores these data chunks and responds to the any chunk requests by providing the chunks. We will see the detailed functioning of these components in the sections below.

A. Cloud Data Distributor

It is an entity between the clients and cloud providers. It receives data from clients in the form of files, slices into small chunks and distributes these chunks to the cloud providers. It also performs receiving chunk requests from clients and forwarding the chunk requests to cloud providers. Clients can interact with cloud providers only through cloud data distributor.

If the client wants to upload any data in cloud, it provides it as files to the cloud data distributor. Clients sends the files with a secrecy level defined for each file. This secrecy level is used to mention sensibility which is a measure used to refer the significance of information that can be seeped through mining data in the file. There are 4 secrecy levels: SL 0, SL 1, SL 2, SL 3.The higher the secrecy level, the data in the file is more sensitive. SL 0 indicates it is the public data(data can be accessible by anyone), SL 1 indicates it is a low sensitive data(data can be retrieved but without the private and protected information), SL 2 indicates it is a moderate sensitive data(protected data used t get any legal information about a company), SL 3 indicates high sensitive data(data used to extract personal information about an individual).

Whenever the cloud data distributor receives a file from client, it partitions the file into several chunks with the same secrecy level for each chunk as the original file. The number of chunks of a file will be intimated to its client by the cloud data distributor, where the clients can request any chunk by mentioning the file name and the serial no of the chunk which corresponds to position of chunk in the file.

Each chunk will be given a unique actual id used to recognize the chunk within cloud data distributor and cloud providers. After this, cloud data distributor distributes the chunks to cloud providers. The cloud provider will not be knowing about owner of the data stored in it. Because the cloud data distributor maintains secrecy level 4 for every cloud provider. Along with the secrecy level, the cloud data distributor also maintains a cost level for every cloud provider. Similar to secrecy level, there are 4 cost levels maintained. Higher the cost level, the cost of storing data in that provider will be high and lower the cost level, the cost of storing data in that provider will be low. The cloud data distributor may also

add some false data into the chunks and these false data can be removed while provided to the clients.

For performing the above mentioned tasks, the cloud data distributor maintains some tables to record the information. They are Cloud Provider Table, Client Table and Data Chunk Table. These tables are described below.

1.Cloud Provider Table: The entries consists about the information of a cloud provider like cloud provider name, its secrecy level SL, cost level CL, no. of chunks given to this provider and list of actual ids of chunks given to this provider.

Cloud Provider Name	SL	CL	No.of Chunks	Actual id list
CLP1	3	2	15289	{24538,…}
CLP2	2	2	60455	{72593,…}
CLP3	2	3	29753	{30674,…}
CLP4	3	3	84670	{92157,…}
CLP5	3	2	56341	{44863,…}

Table 1. Cloud Provider Table

2.Client Table: The entries consists of the information about clients like client name, a pair consisting of password and secrecy level of the password, total no. of chunks of the client and a set of quadruples consisting of filename, serial no., secrecy level and Data Chunk Table index for each chunk belonging to the client.

Client Name	(pw,SL)	No. of Chunks	(filename,sl,SL,cid)
C1	(w31h,2) (59BM,3)	54869	(f1,1,3,0) (f1,2,3,1)
C2	(y4U2,0) (16tG,2)	79043	(f2,1,2,0) (f2,2,2,1)
C3	(al78,3) (8qks,3)	93602	(f3,1,3,0) (f3,2,3,1)

Table 2. Client Table

3.Data Chunk Table:The entries consists of the information about the data chunks like actual id, SL, cloud provider table index of the present cloud provider which is storing the data chunk and set of positions of false data bytes(F) (if any) for all chunks.

Actual id	SL	CLP index	F
24538	3	3	{35,…}
72593	2	1	{72,…}
30674	2	2	{48,…}
92157	3	0	{23,…}
44863	3	1	{28,…}

Table 3. Data Chunk Table

B. Cloud Providers

The next important entity is cloud providers whose responsibilities are to storing data chunks received from cloud data distributor, providing data to clients by replying to a query, and deleting data chunks whenever receiving requests from cloud data distributor..

C. Structural Issues

The above mentioned system structure has several bottlenecks: First the cloud data distributor is the single point of failure. If it crashes or going beyond control, this system will not work. So multiple cloud data distributors can be used to avoid this problem. For each client, one data distributor is primary which is used to upload their data in cloud and other data distributors are secondary which are used to simple data retrieval operations. Second, the reliability of implementing multiple cloud data distributors. They can be implemented at client side itself using hash tables[24] which can map the pair <filename, chunk S1> to a cloud provider. For this, clients have to maintain a data chunk table at their own risk.

D. Simple Application Scenario

Whenever a client wants to run an application using files, the application can request for a single chunk by providing (client name, password, file name, sl.no.) or for all chunks of a file by providing only (client name, password, file name). The password has to be privileged. If the privilege level of the password is greater than or equal to the privilege level of the chunks, the cloud data distributor uses the cloud index field(cid) in the Client Table to identify the respective chunks in Data Chunk Table. The Data Chunk Table issues the Actual id of the respective chunk. This table also provides the cloud provider index(CLP index) which is used to identify the corresponding provider in the Cloud Provider Table. After finding the cloud provider, the cloud data distributor uses the actual id as a key to obtain the necessary chunk from the respective cloud provider. Then the data chunk is passed to the application which is run by the client.

Consider a scenario shown in Figure 3. where a chunk request to cloud data distributor is given as (C1, b25R, f1,0). The client C1 is listed in the Client Table and the password b25R is given in C1 and its secrecy level is 1. The secrecy level of chunk 0 of file 1 is 1. Since the secrecy level of password and chunk are equal, the password will be privileged to request the chunk. Now the chunk index of chunk 0 of file1 is given as 0 in Client Table. So, the Cloud Data Distributor verifies the 0^{th} entry of Data Chunk Table which exposes the actual id of the chunk 11256. And also it provides the present cloud provider index 5 which in turn shows the information about the provider from Cloud Provider Table.

Consider another scenario, where a chunk request to cloud data distributor is given as (C1, 18Ph, f1,0). The client C1 is listed in the Client Table and the password 18Ph is given in C1 and its secrecy level is 0. The secrecy level of chunk 0 of file 1 is 1. Since the secrecy level the password is less than the secrecy level of the chunk, the password will not be privileged to request the chunk and this request will be denied.

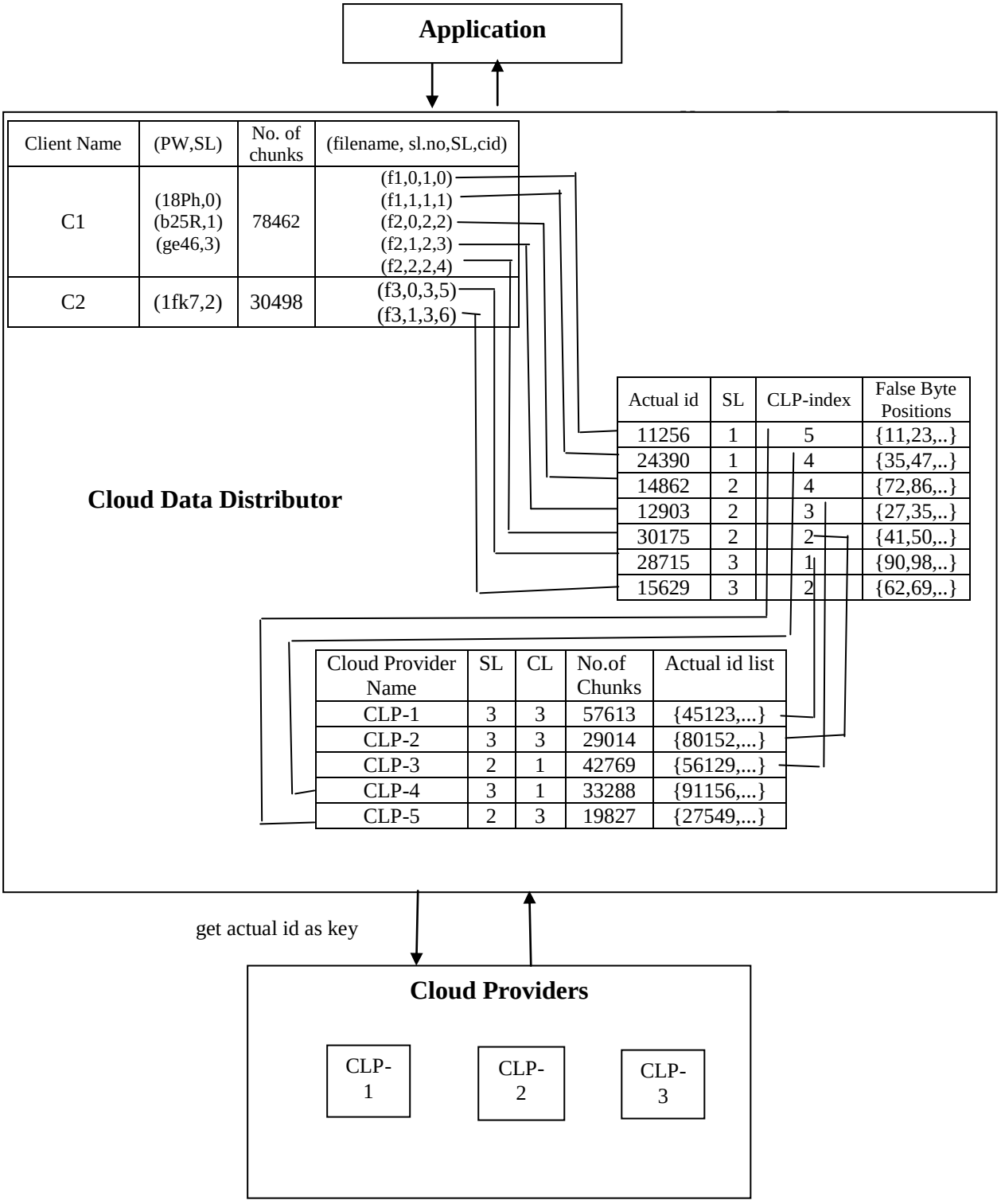

Figure 3. Simple Application Scenario

STEM DESIGN

proposed system, the following functionalities has to be mainly defined.

- Allocate data
- Recover data
- Delete data

The division of data among multiple cloud providers can be implemented by the following 2 functions.

1.datachunks[] divide(file): This function gets a file from client and divides the file into several data chunks. The chunk size will be fixed for a specific privilege level. The higher the privilege level, the lower will be the data chunk size. The actual id attached with each data chunk is used to hide the client identity to the Cloud Provider. Thus the client identity is made private to the Cloud Data Distributor.

2. void allocate(datachunks[]): This function gets chunks from divide[] method and allocates these chunks to the Cloud Providers in a random fashion.

The data recovery process can be implemented by the following 3 functions.

1.datachunk acquire_chunk(clientname,pw,filename,sl.no): This function gets a chunk request from a client, get the required chunk from Cloud Provider and issues it to client.

2.datachunks[] acquire_file(clientname,pw,filename): This function accepts a file request from a client, get all the chunks of the file from Cloud Provider and issues them to client

3.datachunk acquire(actual id as key): This function queries the Cloud Provider for a specific chunk by using the unique actual id as key.

The deletion of data can be implemented by 3 functions.

1.eliminate_datachunk(clientname,pw,filename,sl.no): This function gets a chunk removal request from a client and dispatches the request to the respective Cloud Provider.

2.eliminate_file(clientname,pw,filename): This function gets a file removal request from a client dispatches the request to the respective Cloud Provider.

3.eliminate(actual id as key): This function queries the Cloud Provider to remove a specified chunk by using the actual id as key.

RELATED WORK

In cloud computing, there are various techniques implemented for data assurance. One among the technique is using multiple cloud providers. A similar technique known as RACS(Redundant Array of Cloud Storage) is used nowadays to provide higher assurance of data[4]. Recent related works proposes a new technique Map-Reduce[11] based system which uses Hadoop architecture to provide data privacy and data security in for distributed computation on more sensitive data.

CONCLUSION

There are several type of security threats to the cloud. One among that is a data mining based attack in cloud. In this paper, we have discussed about the siginificance of data mining in cloud and a proposed architecture to avoid data mining attacks which assures the privacy of data in cloud. Our approach uses a multiple cloud provides and a cloud data distributor which performs the categorization of data, fragentation of data into chunks and distribution of the data chunks to multiple cloud providers. In this system, the cloud providers are unaware of the client identity. But if the clients have to restore their data from cloud providers, it will be a diffiult task.

REFERENCES

[1] M.Kantardzic, “Data Mining:Concepts, Models, Methods and Algorithms”, John Wiley & Sons Inc.,2002.

[2] “Introduction to Cloud Computing Architecture”, Sun Microsystems, 2009.

[3] “Top 10 Algorithms in Data Mining”, Springer-Verlag London Ltd.,2007.

[4] H. Abu-Libdeh, L.Princehouse and H.Westherspoon, “RACS:A Case for Cloud Storage Diversity” ACM, pp. 229–240, 2010.

[5] N.Santos,K.P.Gummadi,R.Rodrigues,”Towards Trusted Cloud Computing”,USENIX,2009.

[6] G.Aggarwal, M.Bawa, R.Motwani,”A Distributed Architecture for Secure Databases”, CIDR proceedings, 2005.

[7] Kevin D.Bowers, Ari Juels, Alina Oprea, “HAIL:A High Avaiability and Integrity Layer”, CCS’09,2009.

[8] G.M.Weiss, “Data Mining in the Real World:Experiences, Challenges and Recommendation”, DMIN, pages 124-130, 2009.

[9] Q.Yang, X.Wu, “Ten Challenging Problems in Data Mining Research”, IJITDM,pp 597-604,2006.

[10] R.Chow,P.Golle,M.Jakobsson,E.Shi,J.Staddon,”Controlling data in the Cloud:Outsourcing computation without outsourcing control”, Proceedings of the 2009 ACM Workshop on Cloud Computing Security (CCSW 2009); pp 85-90, 2009.

[11] Jiong Xie,Shu Yin, Zhiyang Ding,”Improving MapReduce Performance through Data Placement in Heterogeneous Clusters”, proceedings in IPDPS,2010.

[12] Jianzong Wang,Zhuo Liu, Peng Wang,”Data Mining of Mass Storage Based on Cloud Computing”.