IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. II (Jan. 2014), PP 66-69
www.iosrjournals.org

Improving Cloud Security Using Data Mining

Srishti Sharma[1], Harshita Mehta[2]
[1](BTech Computer Science and Engineering, VIT University Vellore, India)
[2](BTech Computer Science and Engineering, VIT University Vellore, India)

Abstract: *Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). It does offer great level of flexibility but this advantage comes with a drawback. With increase in sharing of data over web there is an increase in possibility of data being subjected to malicious attacks. Attacker/Provider can extract sensitive information by analyzing the client data over a long period of time. Hence the privacy and security of the user's data is compromised. In this paper we propose an efficient distributed architecture to mitigate the risks.*

Keywords: *Cloud Computing, Distributed Architecture, Malicious Attacks, Security Breaches, Unauthorized access*

I. INTRODUCTION

A cloud is the integration of cloud computing, networking, storage, management solutions, and business applications that aid a new era of IT and consumer services. The services provided by cloud are: infrastructure as a service, platform as a service and software as a service. Cloud services include Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [1]. Many of the major companies like Google, Microsoft and Amazon are providing cloud services. Azure is the cloud computing service provided by Microsoft that will be used by us for implementation [6]. As we move towards an interconnected world of numerous clouds, it gives user access to services anytime and anywhere. But there is also a great risk involved. There are various types of issues that a cloud storage user both at enterprise level and as an individual consumer might face during the use of the service. Most of the issues are with integrity of the data, ensuring that the data is confidential and available when it is needed [3]. Also by monitoring the user data for a long time attacker can extract private information about the user and can use this sensitive data for his/her personal gain. Thus, confidentiality of the data is compromised [2]. Also the user trusting a cloud provider might temporarily or permanently lose the data due to malicious attack. One of the major data analyzing and extracting technique being used now-a-days is data mining. The fact that the entire user data is stored under a single cloud provider gives an opportunity to the attacker to use powerful mining algorithms and extract crucial information about the user [7]. In this paper we solve this issue by dividing the data into chunks after categorizing it and providing these chunks of data to various reliable cloud providers. Categorizing the data helps in determining the sensitive information. Fragmenting and distributing the data among various providers minimizes the information stored under one single provider [2]. Thus the attacker will have access to incomplete information making it difficult for him/her to correlate the data. This optimizes the cost and ensures data privacy but has a significant performance overhead when user has to access all the data very frequently. We propose to store the frequently accessed data separately in the respective providers to reduce space and time overhead. Hence this will prove to be a much more efficient system than the current one.

II. DATA MINING THREATS

Let us consider a scenario where a car Production Company named ABC has trusted a cloud service provider CLD with critical information that includes the history of tender bidding. Now Mr.X who is a malicious employee of company ABC, performed linear multiple regression analysis on the data and found out various dependencies among them. Using these dependencies he calculated the bidding price. Now if X reveals this information to ABC's rival, ABC might lose the next bidding. But if ABC will distribute this data equally among three providers then it will become very cumbersome for the attacker to predict the bidding price, thus saving ABC from a major loss [2].

Suppose a customer shops online from a site. His activity and order summaries are stored in the cloud. Now if anyone has access to this data he/she can analyze the data and find out dependencies among them. These dependencies will help in predicting the client's shopping pattern. Thus, distributing the data will combat this security breach. What we propose is that if the client wants to access his frequent purchases, she/he can easily access them through cache rather than searching the entire data spread across all cloud providers [10].

III. SYSTEM ARCHITECTURE

Cloud Data Distributor and Cloud providers are two major components of the system [1]. The Cloud Data Distributor will receive data in the form of files from the client which will be split into chunks and distributed across various Cloud Providers. The Cloud Provider will store the chunk and after analyzing the data that is frequently accessed by the user he will store this data separately in a different file which will act like the cache. The cloud provider will respond to the queries of the distributor by providing data from cache rather than searching through the entire data chunk which consumes time. Hence, this file containing frequently accessed data will act as a cache memory, increasing the efficiency of the distributed architecture. Clients do not interact with Cloud Providers directly rather via Cloud Data Distributor. This process can be represented as follows:

Client data → Cloud distributor → Data split into chunks and assigned to Cloud Providers → Cloud Provider analyzes the data and stores the frequently accessed data in the cache → Cloud provider responds to the queries of the distributor, depending upon the request from client, through Cache.

The system architecture for a single cloud provider is shown in Fig.1.

IV. IMPLEMENTATION

We have implemented the concept of cache memory in a distributed architecture. Distributed architecture helps in securing user's information. Whenever a user supplies data to a cloud, chunks of data are stored in different cloud providers. When client needs to access all the data frequently, example if client needs to perform global data analysis on all the data, it includes performance overhead. The client may have to access data from multiple locations with a degraded performance. To reduce this overhead we analyze the data. The data is sent to a data mining tool which generates association rules which helps us in determining frequent item sets(having 100% confidence) using Apriori algorithm ,Carma model etc [5].This is called Mining association rules in large transactions and relational databases[4]. Intuitively, a set of items that appears in many baskets is said to be "frequent". To be formal, we assume there is a number s, called the support threshold. If I is a set of items, the support for I is the number of baskets for which I is a subset. We say I is frequent if its support is s or more [8]. Frequent sets of items from data are often presented as a collection of if-then rules. The form of association rules is I→j, where I is a set of items and j is an item. The implication of this associated rule is that if all the items of I appear in some basket, then j is 'likely' to appear in that basket as well. The notion of 'likely' can be formalized by defining confidence of rule I→j to be the ratio of support for I U {j} to the support for I. That is, the confidence of the rule is the fraction of the baskets with all of I that also contain j [8].

These are the steps that were followed to implement this concept:

1. The client data is stored in the database in the cloud (Fig.2) [9].
2. This file is then imported to data mining tool (SPSS MODELER) in excel format. Carma modelling was chosen to generate the association rules and the frequent item sets [11].
3. The data obtained is filtered to obtain the values that have 100% confidence (Fig.3).
4. The data with 100% confidence is then posted on the cloud (Fig.4) [9].

V. FIGURES

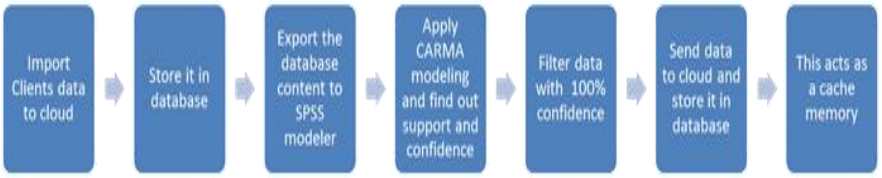

Fig.1 system architecture for a single cloud provider

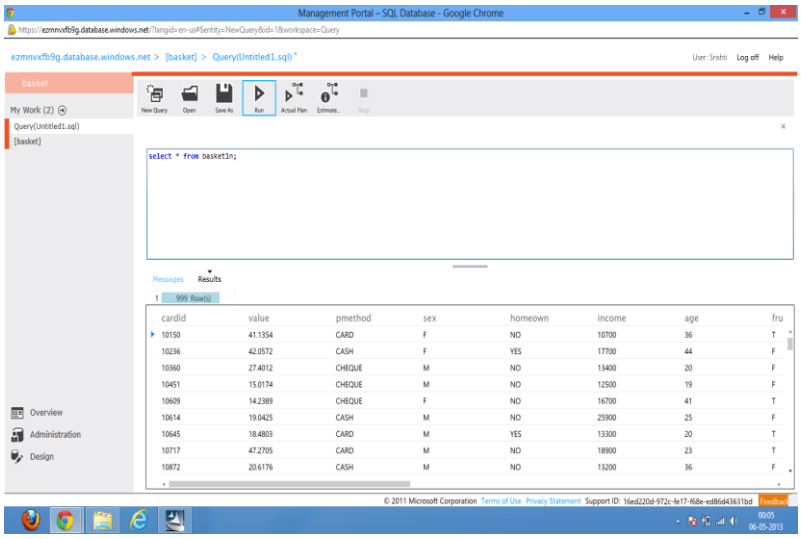

Fig.2 client data stored on the cloud

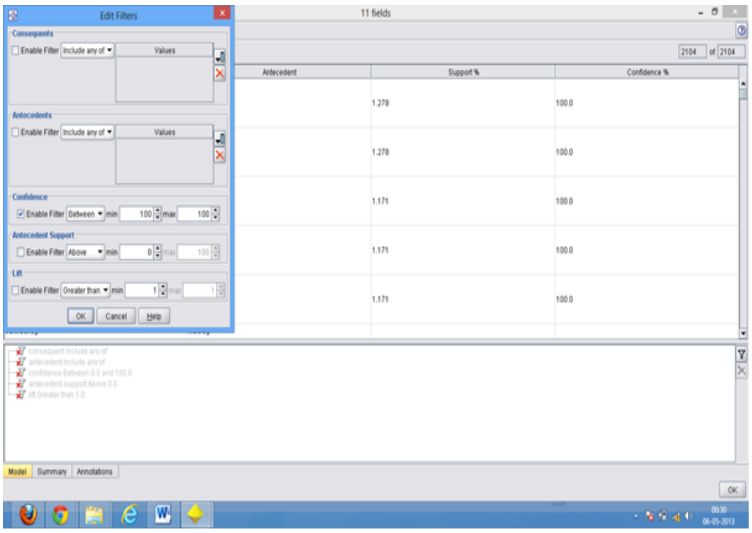

Fig.3 values having 100% confidence are filtered

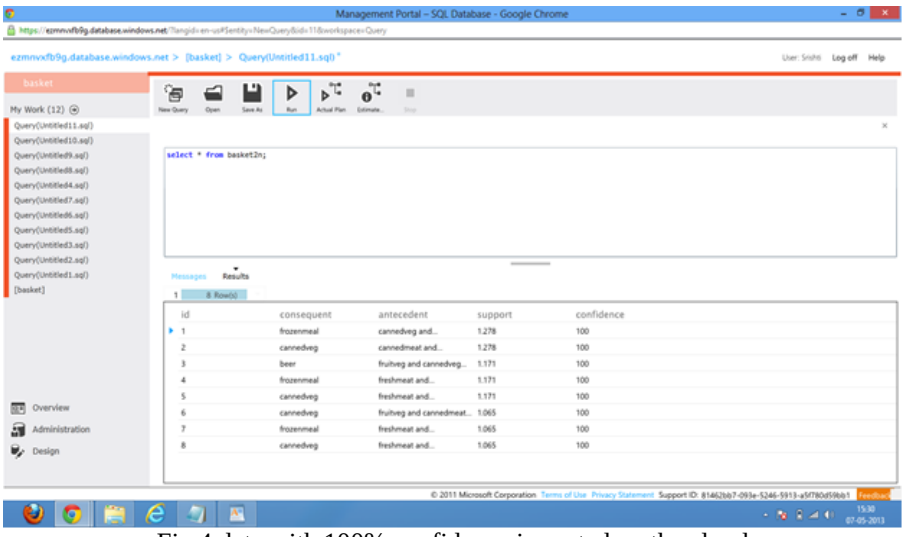

Fig.4 data with 100% confidence is posted on the cloud

VI. Conclusion

Hundred percent secure networks are almost impossible to achieve. New attacks are being discovered every day and new countermeasures have to be developed to keep data secure. Attackers and providers use efficient data mining techniques to extract information about the user from the data stored in cloud. A distributed architecture was proposed to eliminate this threat. But overheads were still prevalent in the system. Hence cache memory concept was implemented in our system by generating frequent item sets using any data mining tool.

In the future instead of having a separate cache for every provider we can have a single cache for all the providers which will store the frequently accessed client data therefore enhancing the efficiency of the current system.

Acknowledgements

We would like to thank Prof. Geetha Mary.A for her support and guidance.

REFERENCES

White Paper:

[1] Introduction to Cloud Computing Architecture by Sun Microsystems,Inc., June 2009

Journal Papers:

[2] Himel Dev, Tanmoy Sen, Madhusudan Basak and Mohammed Eunus Ali, "An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks.", sccompanion, pp.1106-1115, 2012 SC Companion

[3] Anup Mathew, The Institute for Computing, Information and Cognitive Systems (ICICS), University of British Columbia,"Survey Paper on Security & Privacy Issues in Cloud Storage Systems", EECE 571B, TERM SURVEY PAPER, APRIL 2012.

[4] Jiawei Han, "Data Mining Techniques", SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data

Books:

[5] Jiawei Han and Micheline Kamber, Data Mining concepts and Techniques (San Francisco, CA: Morgan Kaufmann, 2006).

[6] William Ryan, Wouter De Kort, Shane Milton , Developing Windows Azure and Web Services, Microsoft Press; 1 edition (14 November 2013)

[7] Siani Pearson, "Taking Account of Privacy when Designing Cloud Computing Services", Software Engineering Challenges of Cloud Computing, 2009. CLOUD '09. ICSE Workshop

Web Sources:

[8] Frequent Itemsets- The Stanford University InfoLab ,http://infolab.stanford.edu/~ullman/mmds/ch6.pdf

[9] Introducing Windows Azure-http://www.windowsazure.com/EN-US/develop/net/fundamentals/intro-to-windows-azure/

[10] J. Salmon, "Clouded in uncertainty – the legal pitfalls of cloud computing", Computing, 24 Sept 2008. http://www.computing.co.uk/computing/features/2226701/clouded-uncertainty-4229153

[11] IBM SPSS Modeler Users Guide ,http://faculty.smu.edu/tfomby/eco5385/data/SPSS/SPSS%20Modeler%2015%20Users%20Guide.pdf