Improving Cloud Security Using Data Mining

Srishti Sharma¹, Harshita Mehta²

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 \rightarrow Cloud distributor \rightarrow Data split into chunks and assigned to Cloud Providers \rightarrow Cloud Provider analyzes the data and stores the frequently accessed data in the cache \rightarrow 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 $I \rightarrow j$, 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 $I \rightarrow j$ 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

FIGURES

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).

V.

4. The data with 100% confidence is then posted on the cloud (Fig.4) [9].



Fig.1 system architecture for a single cloud provider

Management Portal – SQL Database - Google Chrome								- 5	
https://ezmnvxfb9g.database.v	windows.net/?langid=en-us#Sentit	y=NewQuery&id=1&tworkspi	ace=Query						
mnvxfb9g.database.wir	ndows.net > [basket] > Q	(uery(Untitled1.sql)*					User: Srishti	Log off Hel	
basket	പെ								
Work (2) (+)	New Query Open 5	iave As Run Actual Plan	Estimate Stop						
uery((Intitled1.sal)									
asket]									
	Messages Result 1 999 Row(s) Cardid	s value	pmethod	 sex	homeown	income	age	f	
	Messages Result 1 999 Row(s) cardid > 10150	s value 41.1354	pmethod CARD	sex F	homeown	income 10700	age 36		
	Hessages Result 1 999 Row(s) cardid 1 10150 10236	s value 41.1354 42.0572	pmethod CARD CASH	Sex F F	homeown NO YES	income 10700 17700	age 36 44	1	
	Messages Result 1 999 Row(s) cardid > 10150 10236 10960	s value 41.1354 42.0572 27.4012	pmethod CARD CASH CHEQUE	Sex F F M	homeown NO YES NO	income 10700 17700 13400	age 36 44 20		
	Messages Result 1 999 Row(s) Cardid > 10150 10236 10360 10451	s value 41.1354 42.0572 27.4012 15.0174	pmethod CARD CASH CHEQUE CHEQUE	Sex F F M	homeown NO YES NO NO	income 10700 17700 13400 12500	age 36 44 20 19		
	Messages Result 1 999 Rew() Cardid ▶ 10530 10236 10360 10451 10609	s value 41.1354 42.0572 27.4012 15.0174 14.2889	pmethod CARD CASH CHEQUE CHEQUE CHEQUE	Sex F F M M F	homeown NO YES NO NO	income 10700 17700 13400 12500 16700	age 36 44 20 19 41	-	
Overview	Messages Result 1 999 Result cardid ▶ 10150 10236 10460 10451 10609 10514	s value 41.1354 42.0572 27.4012 15.0174 14.2389 19.0425	pmethod CARD CASH CHEQUE CHEQUE CASH	SEX F F M F M F M	homeown NO YES NO NO NO	income 18700 17700 13400 12500 16700 25900	age 36 44 20 19 41 25		
Overview	Messages Result 1 999 Result cardid ▶ 10150 10226 10360 10451 10699 10614 10655	x value 41.1354 42.0572 27.4012 15.0174 14.2389 19.0425 18.4603	pmethod CARD CASH CHEQUE CHEQUE CHEQUE CASH CASH	SEX F F M M F M M	homeown NO VES NO NO NO YES	income 10700 17700 13400 12500 16700 25900 13300	age 36 44 20 19 41 25 20		
Overview Administration	Messages Result 1 999 Result 1 0150 10150 10226 10860 10451 10669 10514 10645 1077	xalue 41.1354 42.0572 27.4012 15.0174 14.2389 19.0425 18.4803 41.2705	pmethod CARD CASH CHIQUE CHEQUE CASH CARD CARD CARD	SEX F M M F M M M M	homeown NO YES NO NO NO NO YES NO	income 10700 117100 113400 12500 16700 25900 13300 18900	age 36 44 19 41 25 20 23	1 7 3 3 3 7 7 7 7	
Overview Administration Design	Heatsages Result 1 999 Board) Cardid 10150 10236 10236 10451 10699 10614 10645 10777 10872	s value 4.1.1554 4.2.0572 27.4012 15.0.174 14.2389 119.045 119.045 18.4603 42.2705 20.6176	pmethod CARD CARD CHEQUE CHEQUE CHEQUE CARD CARD CARD CARD CARD	Sex F M M F M M M M M	homeown NO YES NO NO NO YES NO	income 18700 17700 13400 12500 16700 25900 13300 18900 13200	age 36 44 20 19 41 25 20 23 36	1 7 7 7 7 7 7 7	

Fig.2 client data stored on the cloud

8	Edit Filters	×	11	fields		- 6	
Consequents	(managed and a state of the sta						
Enable Filter	Include any of Values	-1	International Constant State				
		_	Anterlegent	1,278	100.0		
Antecedents	Sectors and w Mature	-					
Conservation	TRAFFIC - TRAFF	×		1.278	100.0		
Confidence	er Detween 🖷 min 🛛 100 🗣 max 🛛 10			1.171	100.0		
Antecedent Sup	er Above Vmin Digman 11			1,171	100.0		
Lit Enable Filter	Oreater than • min 1 2 mar	1		1.171	100.0		
	OK Cancel Help	.,					
월 conseque 월 antereden 월 confidence 월 antereden 월 litt Greater	nt Include any of Enclude any of Exclude any of Exclude 10 and 100 0 Exclude 1 and 100 0 Exclude 1 and 100 0 Exclude 1 and 1 a					X	
Model Summa	ary Annotations					OK.	
0	🕽 🗃 🏉 🗷 🄶				- N 8 -	00.30 06-05-2013	

Fig.3 values having 100% confidence are filtered

🕽 🔠 https://ezmnvafbi9g.database.windov	is.net /Tangidicen-i	s/Sentityn NewC	Ma Svery&id=11&workspi	nagement Portal – SQL Data KerQuery	base - Google Chrom	ne	- 0 <mark>×</mark>
ezmnvxfb9g.database.window	s.net > (baske	t) > Query(U	Intitled11.sq0 *				User: Srishti Log off Help
basket	9a e	a 💾	≥ ⊳∵	- ¹⁷			
My Work (12) 🛞	New Query Ope	n Save.As	Ban Actual Plan	Estimate Stop			
Query(Untitled11.sql)							×
Query(Untitled10.sql)							
Query(Untitled9.sql)	select * fr	om basket2n;					
Query(Untitled8.sql)							
Query(Untitled4.sql)							
Query(Untitled7.sql)							
Query(Untitled6.sql)							
Query(Untitled5.sql)							
Query(Untitled3.sql)							
Query(Untitled2.sql)							
Query(Untitled1.sql)	Messages	Results					
[basket]	1 8	Row(s)					
	id		consequent	antecedent	support	confidence	
	1		frozenmeal	cannedveg and	1.278	100	
	2		cannedveg	cannedmeat and	1.278	100	
	3		beer	fruitiveg and cannedveg	1.171	100	
	4		frozenmeal	freshmeat and	1.171	100	
	5		cannedveg	freshmeat and	1.171	100	
Overview 0	6		cannedveg	fruitweg and cannedmeat	1.065	100	
Administration	7		frozenmeal	freshmeat and	1.065	100	
			canadian	furtherest and	1.065	100	
Design						199	
				© 2011 Mic	rosoft Corporation Terms	of Use Privacy Statement Support ID: 81462bb7-093	e-5246-5913-x5/780d59661 Feedbac
🕑 👩 📋 i	6 🧃	<u>A</u>					→ № 8 ≠ 0 1530 07-05-2013

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
- III]
 IBM
 SPSS
 Modeler
 Users
 Guide

 http://faculty.smu.edu/tfomby/eco5385/data/SPSS/SPSS%20Modeler%2015%20Users%20Guide.pdf
 Guide