Volume 4, Issue 10, October 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Cloud Mining with Privacy Protection

Sakshi Aggarwal[*]
Computer Science Engineering Department
SGT Institute of Engineering & Technology,
Gurgaon, India

Dr. Ritu Sindhu
Associate Professor, CSE Department
SGT Institute of Engineering & Technology,
Gurgaon, India

Abstract— The integration of data mining techniques with cloud computing allows the users to extract the useful information from a data warehouse that reduces the cost of infrastructure and storage. But security and privacy of data is a big concern in data mining on cloud. In current cloud architecture, a client entrusts a single cloud provider with his data. It gives the provider and outside attackers having unauthorized access to cloud, an opportunity of analyzing client data over a long period to extract sensitive information. In this paper, we first identify the data mining based privacy risks on cloud data and provide the information with the help of which data can be secured from unauthorized users.

Keywords— Data Mining, Cloud Computing, Privacy Preserving, RAID, Role Access

I. INTRODUCTION

The importance of Cloud Computing is increasing and it is receiving a growing attention in the scientific and industrial communities. A study by Gartner [1] considered Cloud Computing as the first among the top 10 most important technologies and with a better prospect in successive years by companies and organizations. The Cloud, as it is often referred to, involves using computing resources – hardware and software – that are delivered as a service over the Internet. The use of Cloud Computing is gaining popularity due to its mobility, huge availability and low cost.

Cloud services include Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [2]. Big corporate companies like Amazon, Google and Microsoft are providing cloud services in various forms. Amazon Web Services (AWS) provides cloud services that include Amazon Elastic Compute Cloud (EC2), Simple Queue Service (SQS) and Simple Storage Service (S3) [2]. Although cloud computing is emerging as a technology to achieve high storage and computing services with cost advantages, it brings threats to security of company's data and information. Thus, many potential users and companies lack interest in cloud based services.

Security issues of cloud involve assurance and confidentiality of data. On April 21, 2011, EC2's northern Virginia data centre was affected by an outage and brought several websites down [3]. Problems caused by this outage lasted till April 25,2011 [9]. Such an unlikely event can do significant harm to the users. Confidentiality of user data in the cloud is another big concern. Cloud has been giving providers an opportunity to analyse user data for a long time. In addition, outside attackers who manage to get access to the cloud can also analyse data and violate user privacy. Cloud is not only a source of massive static data, but also a provider of high processing capacity at low cost. This makes cloud more vulnerable as attackers can use the raw processing power of cloud to analyse data. Attackers use the data analysis techniques to extract the valuable information from a large volume of data. These analysis techniques are being used by cloud service providers. For example, Google uses data analysis techniques to analyse user behaviours and recommend search results [4].

Data mining can be a potential threat to cloud security considering the fact that entire data belonging to a particular user is stored in a single cloud provider. The single storage provider approach gives the provider opportunity to use powerful mining algorithms that can extract private information of the user. This approach (single cloud storage provider) also eases the job of attackers who have unauthorized access to the cloud and use data mining to extract information. Thus the privacy of data in the cloud has become a major concern in recent years.

A. Data Mining on Cloud

Data mining is defined as a "type of database analysis that attempts to discover useful patterns or relationships in a group of data. The analysis uses advanced statistical methods, such as cluster analysis, and sometimes employs artificial intelligence or neural network techniques. A major goal of data mining is to discover previously unknown relationships among the data, especially when the data come from different databases." [5]

Data mining techniques and applications are very much needed in the cloud computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining. "Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources.

The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users."The important effect of data mining based cloud computing is that the customer needs to pay only for the data mining tool that he needs. Further the customer need not maintain an infrastructure of as he can use data mining through a browser.

The main effects of data mining tools being delivered by the Cloud are:

- the customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive;
- the customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.

II. SYSTEM THREATS DUE TO CLOUD MINING

Cloud computing needs to address three main security issues:

- Confidentiality
- Integrity
- Availability

Flexibility of services of cloud imposes the risk of security and privacy of user's data. Thus, users of cloud are more concerned about the security of their data when data mining algorithms are used to extract information from data.

There are some problems of data mining based on cloud including:

- The design and selection of data mining algorithms.
- Using appropriate algorithms and adopting appropriate parallel strategy can assist in increasing efficiency.
- Setting appropriate parameters is also very important.
- Privacy protection is a very important issue.

The successful extraction of useful information via data mining depends on two main factors: proper amount of data and suitable mining algorithms. Various mining algorithms are used for numerous purposes. Some mining algorithms are good enough to extract information up to the limit that violates client privacy. Analysis of GPS data is common nowadays and the results of such analysis can be used to create a comprehensive profile of a person covering his financial, health and social status. Thus analysis of data can reveal private information about a user and leaking this sort of information may do significant harm. As more research works are being done on mining, improved algorithms and tools are being developed. Thus, data mining is becoming more powerful and possessing more threat to cloud users. In upcoming days, data mining based privacy attack can be a more regular weapon to be used against cloud users.

A. *Importance of Cloud Privacy*

Client privacy is a tentative issue as all clients do not have the same demands regarding privacy. Some are satisfied with the current policy while others are quite concerned about their privacy. The proposed system is designed preferably for the clients belonging to the second category for which privacy is a great concern.

Data mining is a threat to client privacy. Some mining algorithms allow extracting information up to the limit that violates client privacy. For example, multivariate analysis identifies the relationship among variables and this technique can be used to determine the financial condition of an individual from his buy-sell records, clustering algorithms can be used to categorize people or entities and are suitable for finding behavioural patterns, association rule mining can be used to discover association relationships among large number of business transaction records etc. Thus analysis of data can reveal private information about a user and leaking this sort of information may do significant harm. Especially companies dealing with financial, educational, health or legal issues of people are prominent targets and leaking information of such companies can do significant harm to their customers.

B. *Areas for Secure Cloud Mining Applications*

- Governments can discern illegal or embargoed activities done by individuals, associations or other governments with the implementation of the data mining techniques.
- Businesses can make predictions about how well a product will sell or develop new advertising campaigns by using these new relationships reflected by the data mining algorithms.
- The medical sector benefits from the data mining techniques, as well as the geographical data being better analysed by using data mining.
- In short, data mining has developed uses in the majority of field of activity.

C. *Existing System Threats*

The current cloud storage system is a vulnerable one because data remain under a single cloud provider. This can lead to data loss in case of events like network outage, the cloud provider going out of business, malware attack etc. The current system also gives a great advantage to the attackers as they have fixed targets in the forms of cloud providers. If an attacker chooses to attack a specific client, then he can aim at a fixed cloud provider, try to have access to the client's data and analyse it. This eases the job of the attackers. As long as the entire data belonging to a client remain under a single cloud provider, both inside and outside attackers gets the benefit of using data mining to a great extent. Inside attackers in this context refers to malicious employees at a cloud provider. Data mining models often require large number of observations and single provider architecture is a great advantage suiting the case as all the samples remain under the provider. Thus single provider architecture is the biggest security threat concerning data mining on cloud.

III. A DISTRIBUTED APPROACH TO CLOUD MINING

Data mining on cloud is a potential threat to privacy of user's data because of the fact that entire data belonging to a particular user is stored in single cloud provider. Data mining algorithms require a reasonable amount of data as a result of which the single provider architecture suits the purpose of the attackers.

The job of attackers is also eased because of single cloud storage provider approach. Thus, to protect the privacy of data, approach of distributing the data on cloud to multiple cloud providers can help in protecting the data privacy. The key idea of this approach is to categorize user data, split data into chunks and provide these chunks to the proper cloud providers. This approach consists of categorization, fragmentation and distribution of data.

Categorization allows to identify sensitive data and to take proper initiatives to maintain privacy of such data. Fragmentation and distribution of data among providers reduce the amount of data to a particular provider and thus minimize the risk associated with information leakage by any provider. This distribution is done according to the sensitivity of data and the reliability of cloud providers. A cloud provider is given a particular data chunk only if the provider is reliable enough to store chunks of such sensitivity. Distribution restricts an attacker from having access to a sufficient number of chunks of data and thus prevents successful extraction of valuable information via mining. Even if an attacker manages to access required chunks, mining data from distributed sources remains a challenging job.

Again, mining data from distributed sources is challenging [7]. Specially correlating data from various sources is cumbersome [8] and often leads to unsuccessful mining. So outside attackers managing access to various providers can’t use mining effectively. The distributed approach can take the form of Redundant Array of Independent Disks (RAID) technique used for traditional databases. RACS [4] uses the RAID concept to reduce the cost of maintaining the data on the cloud. It considers each cloud provider as a separate disk. RACS exploits the benefit of RAID on the cloud. For example, RAID level 6 can be used to ensure high assurance of data. It guarantees successful retrieval of data in case of a cloud provider being blocked by any unlikely event or going out of business.

A. System Architecture

This system consists of two major components:

Cloud Data Distributor and Cloud Providers.

The Cloud Data Distributor receives data in the form of files from clients, splits each file into chunks and distributes these chunks among cloud providers.

Cloud Providers store chunks and responds to chunk requests by providing the chunks.

i) Cloud Data Distributor

Cloud Data Distributor receives data (files) from clients, performs fragmentation of data (splits files into chunks) and distributes these fragments (chunks) among Cloud Providers. It also participates in data retrieving procedure by receiving chunk requests from clients and forwarding them to Cloud Providers. Clients do not interact with Cloud Providers directly rather via Cloud Data Distributor. To perform distribution and retrieval of data (chunks), the Cloud Data Distributor needs to maintain information regarding providers, clients and chunks. Hence, it maintains three types of tables describing the providers, the clients and the chunks.

ii) Cloud Providers

The important tasks of Cloud Providers are storing chunks of data, responding to a query by providing the desired data, and removing chunks when asked. Providers receive chunks from the distributor and store them. Each provider is considered as a separate disk storing clients’ data. Certain factors such as distribution of chunks, maintaining privacy levels, reducing chunk size, addition of misleading data contributes to the effectiveness of the system.

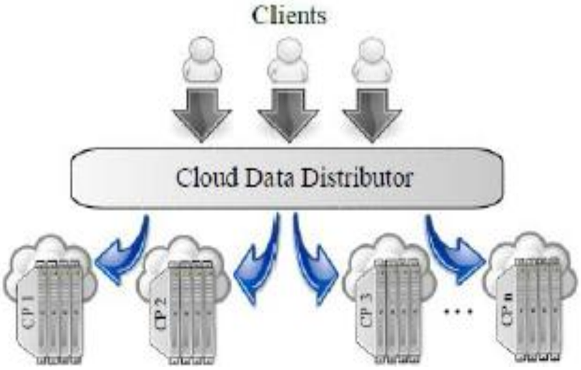

Fig. 1 System Architecture

B. Architectural Issues

The first thing to consider in system architecture is that a single data distributor can create a bottleneck in the system as it can be the single point of failure. To eliminate this, multiple distributors of cloud data can be introduced. In case of multiple data distributors, for each client, a specific distributor will act as the primary distributor that will upload data, whereas other distributors will act as secondary distributors who can perform the data retrieval operations.

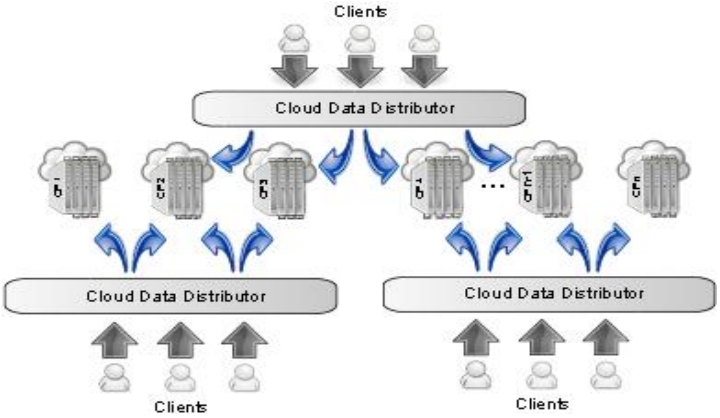

Fig. 2Extended System Architecture

The next issue to consider is the number of privacy levels. Privacy levels in receiving the file chunks from Cloud Data Distributor can be enforced using role hierarchy and password protection.

Whenever the client tries to run any application, application can request for individual chunk by providing the authentication credentials (user name and password).

This privacy level can be increased if the client is assigned a specific role to access the chunk. Roles can be specific to a domain. Roles can be defined according to job competency, authority and responsibility within the enterprise.[8] Thus, file chunks available with cloud distributor are accessible to particular user and having a domain specific role only.

Thus, two level securities are achieved. Level 1 security is enforced by using role based access on file chunks. Unauthorized user will not be able to access the file chunks from Cloud Data Distributor. Level 2 securities are achieved by distributing the file chunks to multiple cloud distributors (i.e. extended architecture).

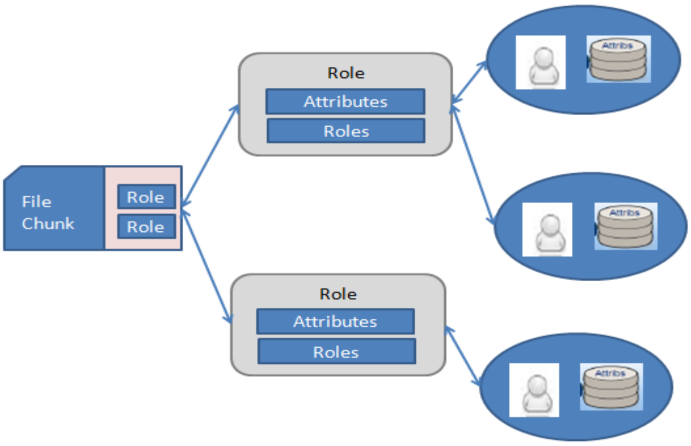

Fig. 3 Role Access Architecture on File Chunks

IV. CONCLUSIONS AND FUTURE WORK

Ensuring cloud data security using data mining on cloud computing is a challenging problem. Various different data mining techniques are used by cloud service providers to extract valuable information from cloud.

In this paper, we have discussed the impact of data mining on cloud computing with the study of distributed cloud architecture to protect the privacy of cloud data. We also proposed the methods to solve the architectural issues in this architecture. The approach extends the distributed cloud architecture for enhanced security and combines the role based security on file chunks in the same architecture.

ACKNOWLEDGMENT

I am thankful to my guide Dr. Ritu Sindhu (Associate Professor, SGT Institute of Engineering & Technology, MDU Rohtak) for taking time to read carefully drafts of this paper and provide the valuable comments. I am also grateful to all the faculty members who has motivated and supported me for preparing this manuscript.

REFERENCES

[1] Gartner Inc *Gartner identifies the Top 10 strategic technologies for 2011*. Online Available: http://www.gartner.com/it/page.jsp?id=1454221. Accessed:15-Jul-2011

[2] M. Brantner, D. Florescu, D. A. Graf, D. Kossmann, and T. Kraska. *Building a database on s3. In J. T.-L.* Wang, editor, ACM, pages 251–264, 2008.

[3] Amazon Web Services: *Overview of Security Processes* may 2011.

[4] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina. *Controlling data in the cloud: Outsourcing computation without outsourcing control*. Pages 85–90, 2009.

[5] Merriam-Webster Dictionary, "*Definition of data mining*", Link: http://www.merriam-webster.com/dictionary/data%20mining.

[6] G. M. Weiss. *Data mining in the real world: Experiences, challenges, and recommendations*. In DMIN, pages 124–130, 2009.

[7] Q. Yang and X. Wu. *10 challenging problems in data mining research*. International Journal of Information Technology and Decision Making, 5(4): 597–604, 2006.

[8] Sunita and Prachi. *Efficient Cloud Mining Using RBAC (Role Based Access Control) Concept*. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013

[9] Wikipedia. *Amazon elastic compute cloud — Wikipedia, the free encyclopedia, 2012*. [Online; accessed 10-May-2011].