

Survey Paper on Data Mining in Cloud Computing

Anuja R. Yeole¹, Poonam Borkar²

¹Student of Master of Engineering in (CSE), Rasoni college of Engineering and Technology, Amravati, India

²Professor, Department (CSE), Rasoni college of Engineering and Technology, Amravati, India

Abstract: In this paper, we describe the privacy of data mining on cloud data that provide the information using which data can be secured from unauthorized users. Data mining is the component which is essential for domain of business. Cloud Computing can make possible model which is good for providers on demand to deploy the computing infrastructure and applications. It gives greater flexibility to allowed access to that enabling services and computing resources. But it has the risk of privacy of data user and security. In recent cloud process, client trust a care of single cloud provider with his data. In today's faster growing IT industry Cloud Computing possesses much more popularity because the Cloud providers believe that it is very easy to handle the data in the cloud ecosystem instead of normal websites in form of simple web pages. Every day the data seeking is being done by indefinite users hugely. Here due to large number of users seeking the data on every day basis there is a serious security interest to the cloud providers as well as the data providers who put their data on the cloud computing ecosystem.

Keywords: Data mining, cloud computing, security, privacy.

1. Introduction

Cloud computing has its own advantages a. It enables end users, small and medium sized users to gain resources which are computational in nature. For example software, storage etc. b. It also allows users to use resources on the cloud and satisfy their necessities. C. A user can access his data anytime anywhere in world without fearing its loss due to system failure or so. Some cloud services are platform as services (PAAS), Infrastructure as service (IAAS), and software as service (SAAS).

- 1) Infrastructure as a Service: It communicates to the PC foundation by using utility administered computer which is regular in a nature. It is also called as utility computation. It gives scalability procedure which is enormous in nature.
- 2) Platform as a Service: It gives approaching towards lease fittings, frameworks which are working, space and system limit that used to create web sites on a top of the infrastructure as a service constructed modeling and combined with middleware ability and improvement and database, informing and ability to hold the queuing.
- 3) Software as a Service: This is a basically clients which are placed to make utilization of an interfaced client to gain other programming have created and invitation as an administration top to the web. It is lying underneath layers of IAAS and PAAS [1].

Famous companies providing cloud services are Google providing Google App Engine, Microsoft SQL Azure, Amazons Elastic Cloud Compute. But at the same time cloud computing is a dual edged sword. It can turn out to be a potential threat to data privacy of user in case the cloud provider fails to secure his cloud services from malware etc. There is also threat of losing the data permanently or temporarily due to miss-handling of cloud provider, which also risks data privacy.

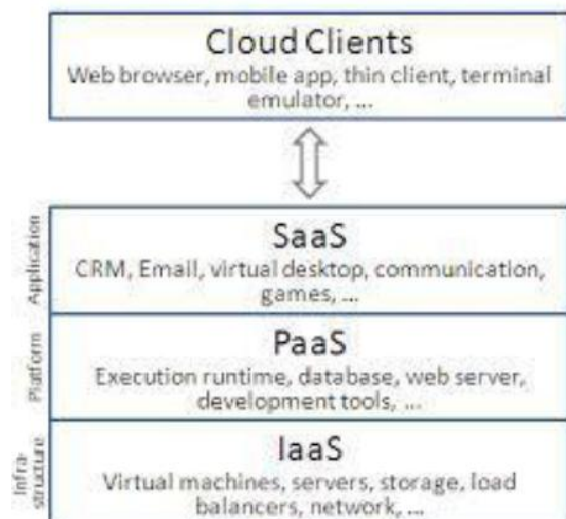


Figure 1: Cloud Computing Layers [2]

1.2 Relation between Data Mining and Cloud Computing

Data mining is a break through invention also known as knowledge Discovery is database [1]. It is used to search significant patterns of data from large volumes of data. Noted areas in Data mining are frequent pattern mining, Association Rule mining etc. Cloud in cloud computing stands for network or Internet or in other words which is present at remote location. It provides services available on WAN, LAN or VPN. For e.g. web conferencing, email etc. Interrelation between data mining and cloud computing have their own pros and cons: Data mining can be used as a tool to provide better services by a cloud provider. But on the other hand outside attackers can use data mining techniques to unauthorizely access private data by interacting it. Interaction of data can involve two factors; one is suitable amount of data and other is appropriate mining algorithms. There are number of mining algorithms which are helpful to interact to private data and hence threat to data privacy. For example association rule mining algorithms can be used to locate relationship between huge numbers of business transaction records.

The three major security issues faced by cloud computing are Confidentiality, Integrity and Availability; respectively. And problems of data mining based on cloud computing.[3]

- a. Design and selection of data mining algorithm.
- b. Using appropriate algorithms and parallel strategy.
- c. Setting of appropriate parameters.
- d. Privacy protection.

A common example of attack on privacy due to data mining is analysis of G.P.S data to create a comprehensive profile of a person covering his financial, health and social status.

- A. Importance of cloud privacy: Sensitivity and awareness may differ from client to client. For some client private is of paramount importance. Certain data mining algorithms threaten client privacy. For example Multi variety analysis identifies relationship between various variables, which in turn can provide financial information of client from his buy sell records etc. similarly clustering algorithms can be used to categorize people or entities to know their behavioral patterns
- B. Area for secure cloud mining Applications: Data mining can apply for various purposes. For example Government can use it for locating illegal or unauthorized activities. Business organizations/Houses can predict outcome of sale of product after a new advertising campaign is launched. Geographical data can better analyze [3].
- C. Existing system threats: Single cloud provider is biggest threat to data privacy. It also eases the job of attacker who just have to target only one specific client. To aggravate the threat to data privacy is involvement of inside attackers i.e employees of cloud provider. Data mining models mostly require larger number of observations and single provider architecture is of great advantage as all the samples remain under provider [4]

2. Literature Review

2.1 Cloud Computing and Its Origin

Internet has been a revolution in the 21st century and with improved bandwidth it has strengthened application services. Cloud computing serves as a platform for converting information technology resources into services for end users, software services, basic infrastructure services etc. In cloud computing user need not install any software on his side but may use all the hardware resources in cloud as client side infrastructure. Cloud computing like any other service is client or demand driver. Cloud computing defined as a type which is distributed system and parallel set of items which are inter connected and computer virtualized that are provisioned to be dynamic as one or more single entity of computing resources based on level of agreements which is established through part of negotiation between service providers and consumers".[5]

To build cloud computing application as service, infrastructure, platform and application software is required. These may be obtained from single or several service providers. If income for cloud service comes from charging for infrastructure than it is called Infrastructure as service, if income comes platform than "Platform as service" and if it comes from application software than the business model is

software application as service. It builds user interface and portrays application system's functions.

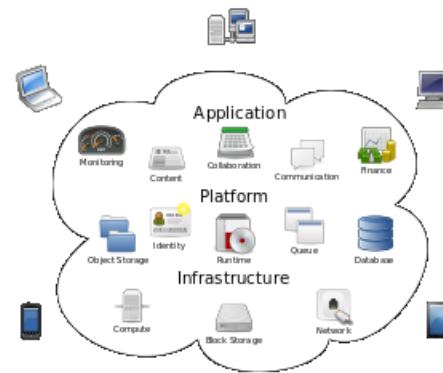


Figure 3.1: Cloud computing

2.2 Data Mining and its Origin

Data mining prospered in 1990's but has a long history behind its evolution. History of data mining can be traced in classical statistics, artificial intelligence and machine learning. Statistics includes standard distribution, standard derivation, standard variance, cluster analysis, discriminate analysis etc. All these play role in analyzing data and data relationships. Artificial intelligence applies to processing human thought to statistical problems. Certain commercial products like RDMS have used query optimization modules which is a concept of Artificial Intelligence.

Another concept of machine learning allows program to study data and then make decisions on the basis of data studied. Programmers uses statistics for fundamental concepts adding advanced AI heuristics, algorithms for the above given purpose. Thus data mining is nothing but application of machine learning techniques to business applications.

Techniques like AI, statistics, machine learning are used to find previously hidden trends or practices or data. It is nothing but data mining. Association Rule which finds association between data and various objects by finding dependence amongst data is often used method in data mining. One thing has to be discussed here that business must hide complexity of data mining from end-user to make a successful application.[6]

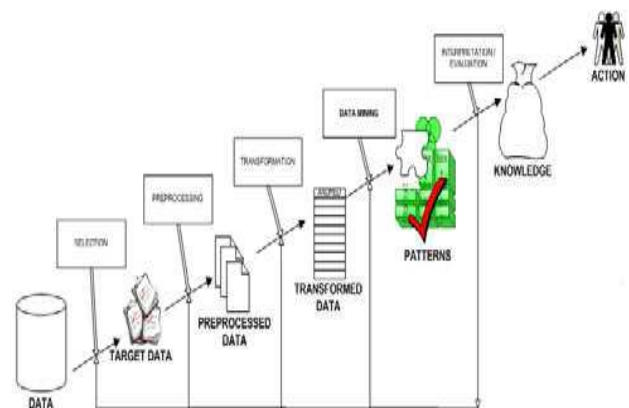


Figure 3.2: Steps in Data Mining Process

3. Proposed Work

3.1 Preserving Privacy from Data Mining Cloud

3.1.1 Protection from Data Mining Based Attack Using Distributed Cloud Architecture

These days outside attackers or cloud provider can use data at analysis techniques to entrant sensitive personal information from cloud. One of them is data mining algorithm aimed at single cloud. The present paper attempts to provide solution to this problem by providing unique identify to cloud users and servers such as Federal Identify Management multiple cloud providers can prevent data mining techniques.

Lets first discuss what is Multiple cloud provider. In this process user data is in the smaller parts or section and then is sent to multiple cloud providers. The process involves there steps

1st step- classification: In this step sensitive private data is identified and requisite mechanisms are applied to secure data privacy of user.

2nd step- Fragmentation: After identification of private data it is further divided into small sections/compartments.

3rd step- Distribution: After bifurcation the small parts of data are sent to cloud providers depending upon their reliability. A cloud provider is reliable if his capable of securing private data.

This technique makes the job of attackers impossible; as it is very difficult to collect smaller parts of data from different cloud providers.

The above given system consists of two major components i.e. cloud data distributor and cloud provider.

1) **Cloud Data Distributor**- It receives data, divides it in to fragments or sections and distributes it among various cloud providers. Client is not aware of the data providers. Cloud data distributor keeps three types of information one is fragmented data, second is client and third is cloud provider.

2) **Cloud provider**- It receives fragmented data which it preserves. It also responds to query concerning data when inquired about and removes the same when required so. The system becomes more effective by distribution of fragments/sections of data maintaining privacy levels reducing section's size, addition of misleading data etc.

3.1.2 Cloud Computing Security

Data in cloud can be effectively secured by encrypting it. Direct access of client can be restricted by using proxy and brokerage services.

Security Planning- If a resource is to be deployed to cloud than several attributes about the resource should be analyzed [2].

- 1) A resource's sensitivity to risk is to be first analyzed.
- 2) Various cloud service models like IAAS, PAAS or SAAS are to be considered which requires consumers to be responsible for security at different levels.
- 3) Different types of clouds should be considered for example Public, Private, Community or Hybrid.

4) Cloud service providers system is to be analyzed as to how data is transferred, stored or moved in to or out of cloud [7] .

4. Conclusion

Privacy is most sacred right and expectation of a human being which has to be preserved. It is important to analyze result of these researches and study security requirements to keep data private and secure. Cloud computing has its advantages in a sense that end user need not invest on infrastructure. But it has to be kept in mind that it is fragile to data mining techniques used by attackers who gain unauthorized access, risking data privacy. Hence there is need to evaluate security measures and protect client privacy.

References

- [1] D. Bhu Lakshmi and S. Arundathi, "Providing Privacy and Security for Cloud Data Using Data Mining," International Journal of Innovation and Scientific Research (Vol. 11 No. 2, Nov 2014) ISSN 2351-8014., pp. 264-272 Available at <http://www.ijsr.issr-journals.org/>
- [2] Inderjit Kaur, Deep Mann, "Data Mining in Cloud Computing International Journal of Advanced Research in Computer Science and Software Engineering" (Volume 4, Issue 3, March 2014) ISSN: 2277 128X, available at http://www.ijarcsse.com/docs/papers/Volume_4/3_March_2014/V4I3-0601.pdf
- [3] K Sasirekha, Dr K. Raja, "An Approach To Improve Cloud Data Privacy By Preventing From Data Mining Based Attacks," International Journal of Scientific and Research Publications, (Volume 4, Issue 2, February 2014) ISSN 2250-3153 available at <http://www.ijarp.org/research-paper-0214/ijarp-p2698.pdf>
- [4] Sakshi Agarwal, Dr Rithu Sindhu, "A Survey on Cloud Mining with Privacy Protection" International Journal Of Engineering And Computer Science (Volume 2 Issue 3 March 2013) ISSN:2319-7242, Pp. 545-550 available at http://www.ijarcsse.com/docs/papers/Volume_4/10_October2014/V4I10-0165.pdf
- [5] Cloud Computing (Feb 2014), available at http://en.wikipedia.org/wiki/Cloud_computing
- [6] Infra 1
- [7] Himel Dev, Tanmoy Sen, Madhusudan Basak and Mohammed Eunus Ali, "An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks" Department of Computer Science and Engineering (CSE) Bangladesh University of Engineering and Technology (BUET) available at http://www.cse.buffalo.edu/faculty/tko/sa/dacloud2012/papers/datacloud2012_paper_8.pdf