International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 11 No. 2 Nov. 2014, pp. 264-272

http://www.ijisr.issr-journals.org/

Providing Privacy and Security for Cloud Data Using Data Mining

D. Bhu Lakshmi[1] and S. Arundathi[2]

[1]Department of Computer science and engineering,
KNS Institute of technology,
Bangalore, Karnataka, India

[2]Department of Master of Computer Applications,
KNS Institute of technology,
Bangalore, Karnataka, India

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. In this paper, review of various data mining techniques is presented which can help achieve security of information on cloud. In today's fastest growing IT industry Cloud Computing is gaining much more popularity because the Cloud providers feel that it is very easy to manage the data in the cloud environment rather than normal web-sites in form of simple web pages. Every day the data seeking is being done by many users immensely. Here due to immense number of users seeking the data on daily basis there is a serious security concerns to the cloud providers as well as the data providers who put their data on the cloud computing environment. This paper provides solution of the security issues of cloud computing**.**

KEYWORDS: Cloud Computing, Data Mining, Security, Privacy etc.

1 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 loose 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 maindrawback 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].

2 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.

The term Cloud refers to a Network or Internet. In other words, we can say that Cloud is something which is present at remote location. Cloud can provide services over network i.e. on public networks or on private networks i.e. WAN, LAN or VPN. Applications such ase-mail, web conferencing, customer relationship management (CRM), all run in cloud.

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.

CLOUD COMPUTING SECURITY CHALLENGES

Data protection tops the list of cloud concerns today. Vendor security capabilities are key to establishing strategic value, reports the 2012 Computerworld “Cloud Computing” study, which measured cloud computing trends among technology decision makers.

When it comes to public, private, and hybrid cloud solutions, the possibility of compromised information creates tremendous angst. Organizations expect third-party providers to manage the cloud infrastructure, but are often uneasy about granting them visibility into sensitive data.

There are complex data security challenges in the cloud:

- The need to protect confidential business, government, or regulatory data
- Cloud service models with multiple tenants sharing the same infrastructure
- Data mobility and legal issues relative to such government rules as the EU Data Privacy Directive
- Lack of standards about how cloud service providers securely recycle disk space and erase existing data
- Auditing, reporting, and compliance concerns
- Loss of visibility to key security and operational intelligence that no longer is available to feed enterprise IT security intelligence and risk management

Specific security challenges pertain to each of the three cloud service models:

Software as a Service (SaaS)
Platform as a Service (PaaS)
Infrastructure as a Service (IaaS)

- **SaaS** deploys the provider's applications running on a cloud infrastructure; it offers anywhere access, but also increases security risk. With this service model it's essential to implement policies for identity management and access control to applications. For example, with Salesforce.com, only certain salespeople may be authorized to access and download confidential customer sales information.
- **PaaS** is a shared development environment, such as Microsoft™ Windows Azure, where the consumer controls deployed applications but does not manage the underlying cloud infrastructure. This cloud service model requires strong authentication to identify users, an audit trail, and the ability to support compliance regulations and privacy mandates.
- **IaaS** lets the consumer provision processing, storage, networks, and other fundamental computing resources and controls operating systems, storage, and deployed applications. As with Amazon Elastic Compute Cloud (EC2), the consumer does not manage or control the underlying cloud infrastructure. Data security is typically a shared responsibility between the cloud service provider and the cloud consumer. Data encryption without the need to modify applications is a key requirement in this environment to remove the custodial risk of IaaS infrastructure personnel accessing sensitive data.

TECHNIQUES FOR PROTECTING DATA IN THE CLOUD

Traditional models of data protection have often focused on network-centric and perimeter security, frequently with devices such as firewalls and intrusion detection systems. But this approach does not provide sufficient protection against APTs, privileged users, or other insidious types of security attacks.

Many enterprises use database audit and protection (DAP) and Security Information and Event Management (SIEM) solutions to gather together information about what is happening. But monitoring and event correlation alone do not translate into data security.

At a time when regulation and compliance issues are at an all-time high, it's dangerous to assume that monitoring, collecting, and storing logs can protect the organization from security threats, as they are reactive controls. In today's environment, both data firewalls and data security intelligence are essential to adequately protect the enterprise from new and different types of attacks.

It's critical that CISOs implement a data security strategy that provides a veritable firewall around the data itself for comprehensive protection. Advanced data security solutions provide CISOs with an early warning system about an attack, render the content unusable, and leverage automation and big data analytics to continuously analyse logs and other information about their environment such as security events and data flow.

While many organizations have implemented encryption for data security, they often overlook inherent weaknesses in key management, access control, and monitoring of data access. If encryption keys are not sufficiently protected, they are vulnerable to theft by malicious hackers. Vulnerability also lies in the access control model; thus, if keys are appropriately protected but access is not sufficiently controlled or robust, malicious or compromised personnel can attempt to access sensitive data by assuming the identity of an authorized user.

2.1 ORIGIN AND DEFINITION OF CLOUD COMPUTING

The Internet rapidly began to grow up in the 1990s and, the progressively more complicated network infrastructure and enlarged bandwidth developed in the recent years have considerably improved the strength of various application services available to users through the Internet, hence, marking the beginning of cloud computing network services. Cloud computing services use the Internet as a communication medium and convert information technology resources into services for end-users, including software services, computing platform services, development platform services, and basic infrastructure leasing.

Primary significance of Cloud computing lies in allowing the end users to access computation resources through the Internet. The unusual features of cloud computing include the storage of user data in the cloud and the lack of any need for software installation on the client side. Provided that the user is able to connect to the Internet, all of the hardware resources in the cloud can be used as client-side infrastructure. Normally, cloud computing applications are demand-driven, providing various services according to user requirements, and service providers charge by metered time, instances of use, or

defined period. Cloud computing can be defined as "a type of parallel and distributed system which consists of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers".[1] The cloud computing concept can be understood in a more better way by following the below given figure:[1]

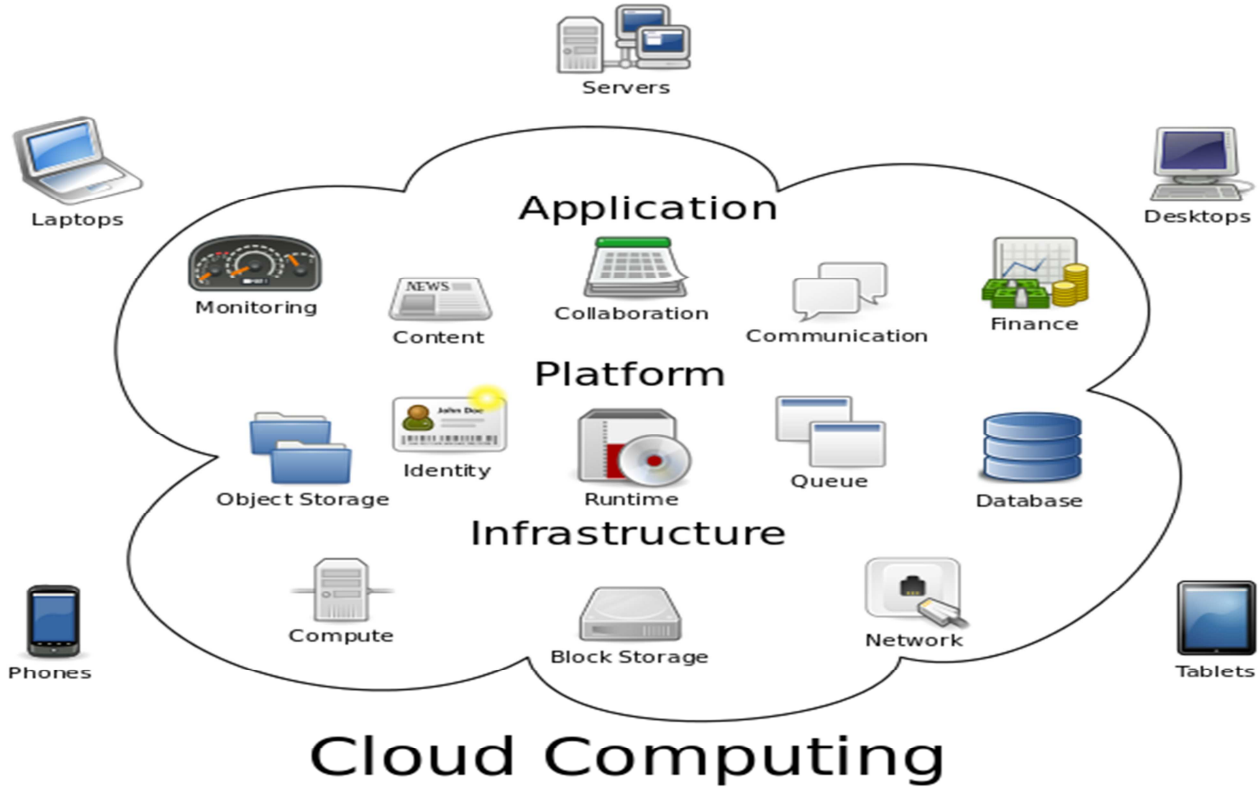


Fig. 1. Figure to represent cloud computing concept map

The architecture of cloud services can be divided into three levels: infrastructure, platform, and application software. Application software builds the user interface and shows the application system's functions. To build a cloud computing application as a service requires infrastructure, platform and application software which can be obtained from a single provider or from different service providers. If the income for cloud services mainly comes from charging for infrastructure, this business model can be referred to as Infrastructure as a Service (IaaS). If income comes mainly from charging for the platform,the business model can be referred to as Platform as a Service (PaaS). If income mainly comes from charging for applications or an operating system, the business model can be referred to as Software as a Service (SaaS). The model being proposed in this paper uses SaaS concept.

2.2 ORIGIN AND DEFINITION OF DATA MINING

Data mining is to find knowledge, and knowledge is represented through certain patterns. Association rule is the most often used method in data mining, which finds out the association between data and various objects by finding the potential dependence among data. Classification and clustering can be used to sort out things by characterizing the common significance among different things. The disadvantage of data mining in centralized database, generally have the several following points: network traffic is considered less, mining efficiency is low and the degree of spatial complexity is high. The most classic classification data mining are classification methods based on distance, classification methods based on decision tree, Bayesian classification and so on. Data mining techniques have been extensively used in various applications. However, the mistreat of these techniques may lead to the discovery of sensitive information. Researchers have recently made efforts at hiding sensitive association rules. However, undesired side effects, e.g., non-sensitive rules falsely hidden and spurious

rules falsely generated, may be formed in the rule hiding process. [5] Privacy has become an significant issue in Data Mining. Many methods have been brought out to solve this problem. The basic aspect which we are concerned about in this paper is of association rule mining which preserves the confidentiality of each database. In order to find the association rule, each participant has to share their own data. Thus, a lot of privacy information may be put out or been illegally used. [6] Data mining can be defined as "the process that attempts to discover patterns in large data sets". The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The following figure explains the different steps which comprise the overall data mining process:

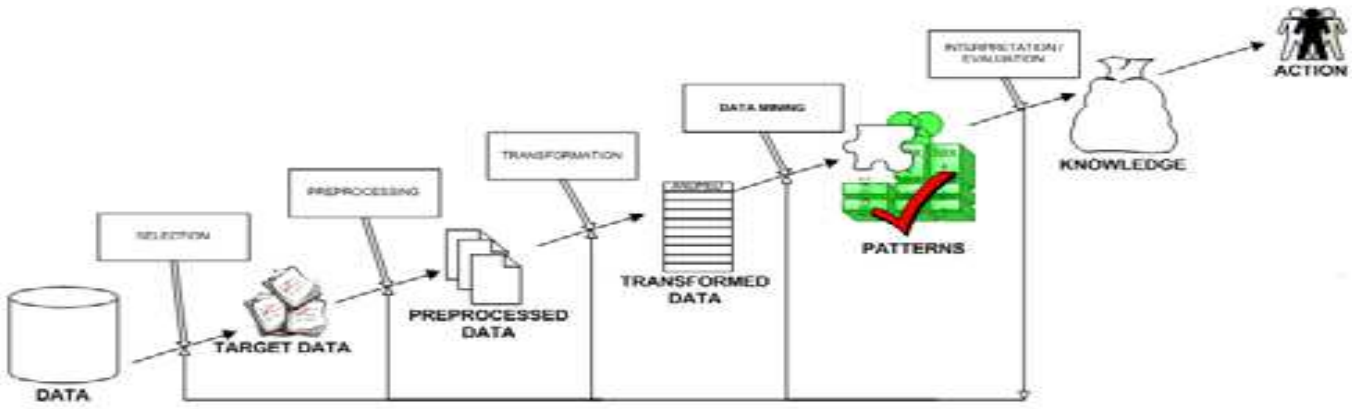


Fig. 2. Figure to represent steps in data mining process

3 PRIVACY PRESERVING DATA MINING IN CLOUD

3.1 PROTECTION FROM DATA MINING BASED ATTACK USING DISTRIBUTED CLOUD ARCHITECTURE

Data mining can be a potential threat to cloud security because of the fact that entire data belonging to a particular user is stored in a single cloud provider. The provider gets an opportunity due to a single storage provider approach to use powerful mining algorithms or tools that can extract private information of the user. 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. These attackers have unauthorized access to the cloud and use data mining to extract information.

In this approach data is distributed multiple cloud providers so that data mining becomes a difficult job to the attackers. 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. The categorization of data is done according to mining sensitivity. Mining sensitivity in this context refers to the significance of information that can be leaked by mining. 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.

This distributed approach provides two major benefits first, it improves privacy by making the attacker's job complicated by increasing the number of targets and decreasing amount of data available at each target. Second, it ensures the greater availability of data.

This system consists of two major components Cloud Data Distributor and Cloud Providers[14]. 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.

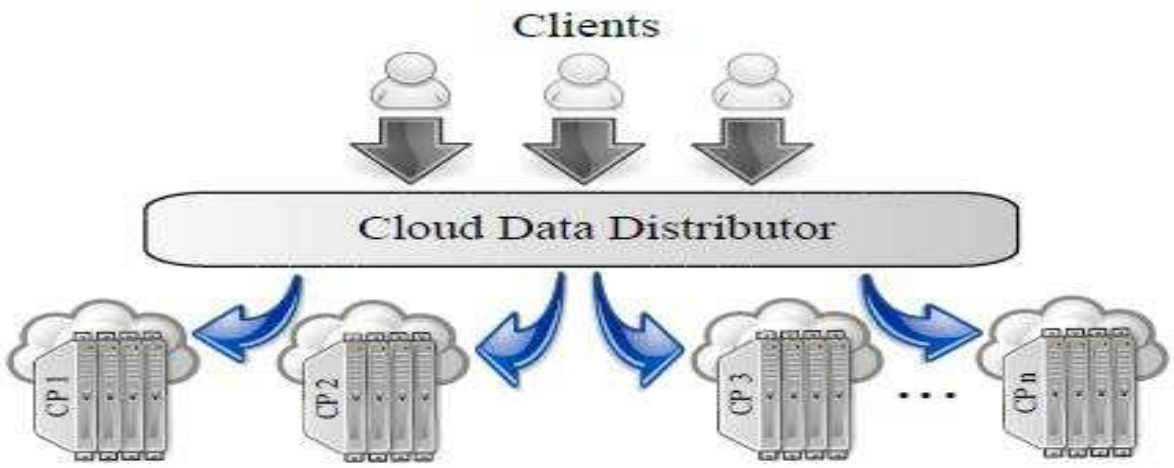


Fig. 3. System Architecture

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

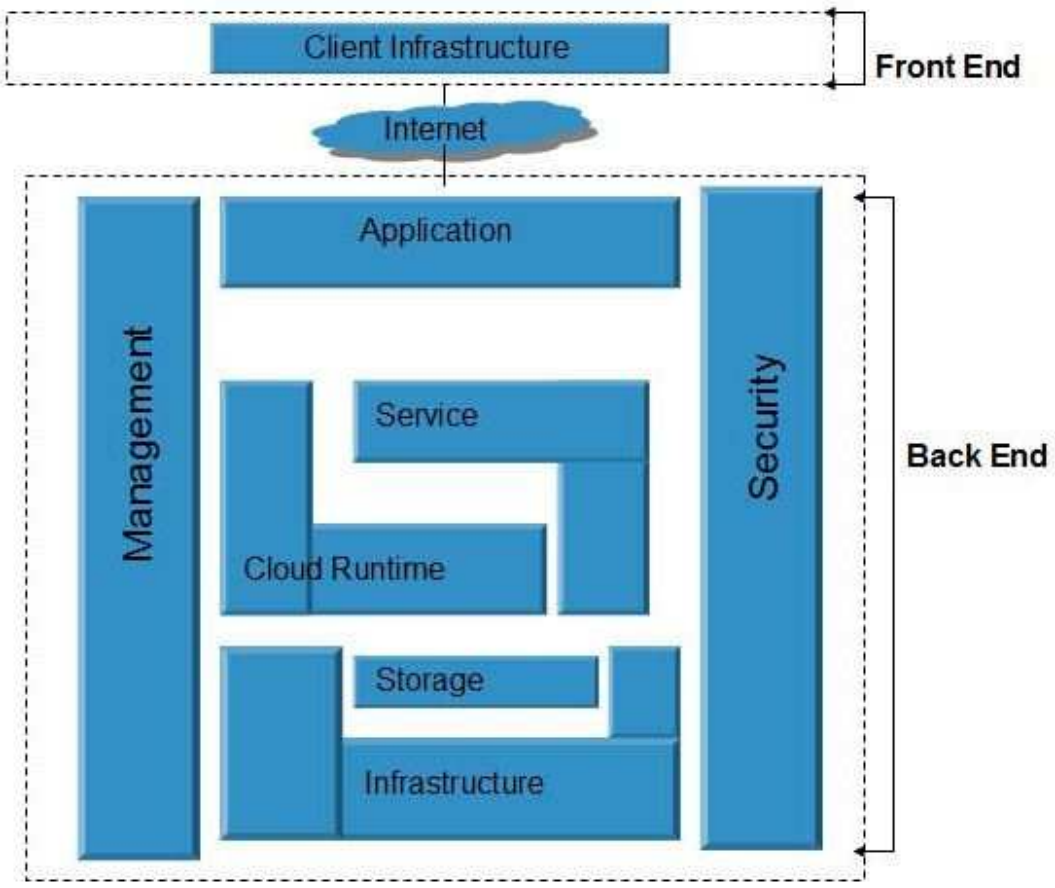
The important tasks of Cloud Providers are storing chunks of data, responding to a query by providing the desired data, and removing chunks when asked. Providers receive chunks from the distributor and store them. Each provider is considered as a separate disk storing clients' data. Certain factors such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

4 CLOUD ARCHITECTURE

The Cloud Computing architecture comprises of many cloud components, each of them are loosely coupled. We can broadly divide the cloud architecture into two parts:

- Front End
- Back End

Each of the ends are connected through a network, usually via Internet. The following diagram shows the graphical view of cloud computing architecture:



FRONT END: refers to the client part of cloud computing system. It consists of interfaces and applications that are required to access the cloud computing platforms, e.g., Web Browser.

BACK END: refers to the cloud itself. It consists of all the resources required to provide cloud computing services. It comprises of huge data storage, virtual machines, security mechanism, services, deployment models, servers, etc.

5 CLOUD COMPUTING SECURITY

Security in cloud computing is a major concern. Data in cloud should be stored in encrypted form. To restrict client from direct accessing the shared data, proxy and brokerage services should be employed.

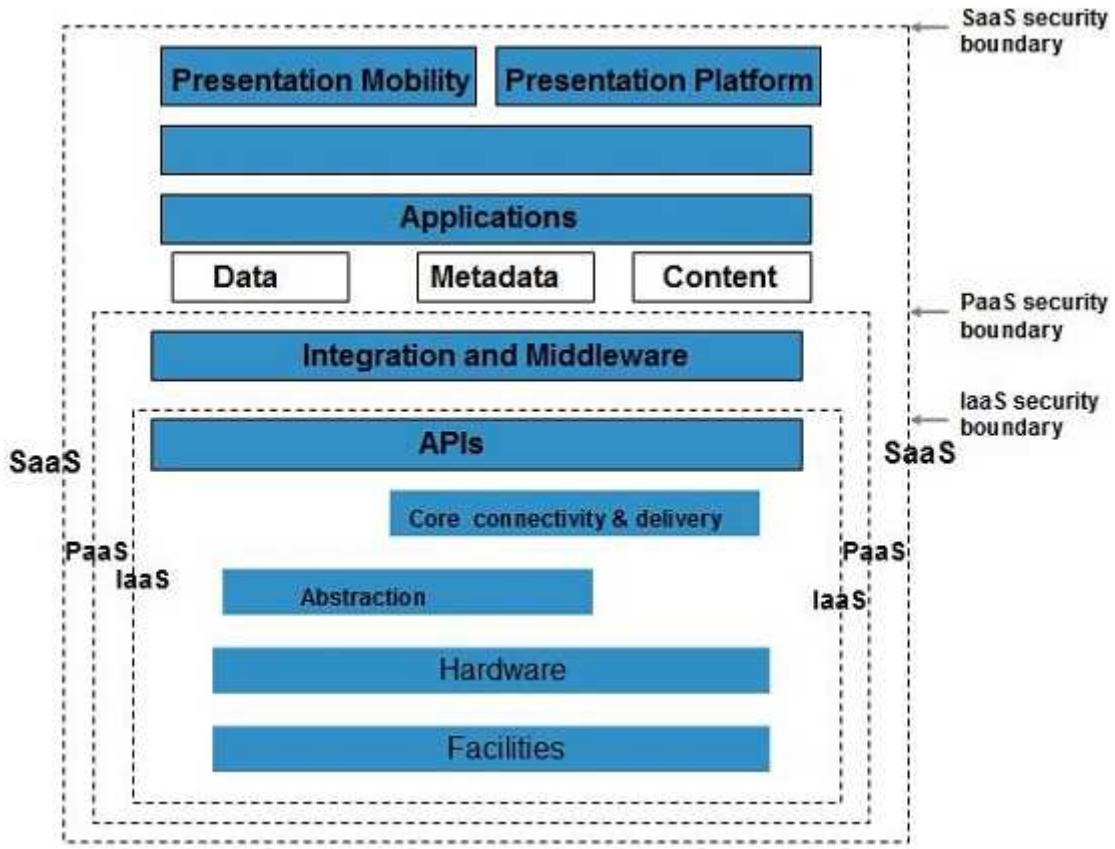
SECURITY PLANNING

Before deploying a particular resource to cloud, one should need to analyze several attributes about the resource such as:

- Select which resources he is going to move to cloud and analyze its sensitivity to risk.
- Consider cloud service models such as **IaaS, PaaS**, and **SaaS**. These models require consumer to be responsible for security at different levels of service.
- Consider which cloud type such as **public, private, community** or **hybrid**.
- Understand the cloud service provider's system that how data is transferred, where it is stored and how to move data into and out of cloud.

SECURITY BOUNDARIES

A particular service model defines the boundary between the responsibilities of service provider and consumer. **Cloud Security Alliance (CSA)** stack model defines the boundaries between each service model and shows how different functional units relate to each other. The following diagram shows the **CSA stack model:**



KEY POINTS TO CSA MODEL

- IaaS is the most basic level of service with PaaS and SaaS next two above levels of service.
- Moving upwards each of the service inherits capabilities and security concerns of the model beneath.
- IaaS provides the infrastructure, PaaS provides platform development environment and SaaS provides operating environment.
- IaaS has the least level of integrated functionalities and integrated security while SaaS has the most.
- This model describes the security boundaries at which cloud service provider's responsibility ends and the consumer's responsibilities begin.
- Any security mechanism below the security boundary must be built into the system and above should me maintained by the consumer.
- Although each service model has security mechanism but security needs also depends upon where these services are located, in private, public, hybrid or community cloud.

6 CONCLUSION

Cloud service providers as well as other third parties use different data mining techniques to acquire valuable information from user data hosted on the cloud. We have discussed the impact of data mining on a single cloud and have proposed a distributed structure to eliminate mining based privacy threat on cloud data. Finally, we also discussed a technique to secure or protect the privacy of the forecasting reports for the company. These reports could be used by companies to increase their sales. This technique protects the predictions that are generated as a result of mining and secure it from interception.

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
[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 Map Reduce 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".
[13] Himel Dev, Tanmoy Sen, Madhusudan Basak, and Mohammed Eunus Ali, "An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks",. 2012 SC Companion: High Performance Computing, Networking Storage and Analysis.
[14] Introduction to Cloud Computing Architecture by Sun Microsystems, Inc., june 2009.
[15] Amazon Web Services: Overview of Security Processes, may 2011.
[16] H. Abu-Libdeh, L. Princehouse, and H. Weatherspoon. Racs: a case forcloud storage diversity. In ACM SoCC, pages 229–240, 2010.
[17] G. Aggarwal, M. Bawa, P. Ganesan, H. Garcia-molina, K. Kenthapadi, R. Motwani, U. Srivastava, D. Thomas, and Y. Xu. Two can keep a secret: A distributed architecture for secure database services. In InProc. CIDR, 2005.
[18] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. Above the clouds: A berkeley view of cloud computing. Technical report, EECS Department, University of California, Berkeley, 2009.
[19] P. Asst. Prof . PSV Vachaspati. Quantum attack resistant cloud. World of Computer Science and Information Technology Journal, 1:283–288,2011.
[20] M. Bramer. Principles of Data Mining. Springer, 2007.
[21] 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.
[22] S. H. Brown. Multiple linear regression analysis: A matrix approach with matlab. Alabama Journal of Mathematics, 2009.
[23] 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.
[24] C. Clifton and D. Marks. Security and privacy implications of datamining. In ACM SIGMOD Workshop, pages 15–19, 1996.