Data Replication System in Secure Data Access control in Cloud Environment

1 Vellanki Vijayababu, 2P. Nageswara Rao, 3S. Suresh
1,2,3 Dept. of CSE, Sree Rama Institute of Technology & Science, SKuppanakuntla, Penuballi, Khammam, AP, India

Abstract: Replication is a widely used method for achieving high availability in database systems. The data replication approach can be used for solutions such as Sharing data among remote offices, Sharing data among dispersed users, Make server data more accessible, Distribute solution updates, Back up data, Provide Internet or Intranet Replication. Cloud computing is most prevalent technology used in various organizations. It provides wide range of applications that are useful for various business needs. It provides the facility to store and access the data from any place around the world across distributed network environment. Data stored in the network is always vulnerable to threat from an outsider and as well as from a malicious insider. Various encryption mechanisms have been used to provide data security. But they have failed to provide enough protection to the data. We propose a different approach for providing security to the data in the cloud by using deception technique. We monitor the different user’s search behaviour to analyze their search patterns. Using these different patterns we detect abnormal data access patterns. Once illegal data access is identified and verified by answering security question, we launch disinformation attack and ensure to minimize data theft and misuse of user’s real data. Keywords: Data access, Data security, malicious insider, decoy information, SVM, multi clouds, big data analytics.

Introduction

Database Replication is the frequent electronic copying of data from a database in one computer or server to a database in another so that all users share the same level of information. The result is a distributed database in which users can access data relevant to their tasks without interfering with the work of others. Cloud computing provides users with flexible, Cost Efficient, limited Storage, Backup, Recovery, Easy Access of Information and provides better Deployment facilities to the user’s of the cloud. Businesses like small, medium, and large companies prefer to outsource their services to cloud for better operational efficiency. Cloud is just not a commodity plan for cutting the cost of running a few applications, but it is also about transforming business. Cloud has 3 service models IaaS, PaaS, SaaS. The bottom layer is Infrastructure as a Service (IaaS) provides virtual machines and other resources like block and files storage, network Security, load balancing, virtual local area networks (VLANs) etc. IaaS cloud providers supply these resources on-demand from their large pools installed in their centres. For wide-area connectivity, customers can use either the Internet or carrier clouds (i.e. dedicated virtual private networks). The second layer from the bottom is Platform as a Service (PaaS) here, cloud service providers deliver a computing platform like operating system, execution environment (programming language), database, and web servers. Application developers can develop and run their software on a cloud platform without buying and managing the hardware and software layers. Some PaaS service providers like Windows Azure, Google AppEngine enable the computers and storage resources vary automatically to match application demand so that the cloud user does not have to allocate resources...
The last service model is Software as a Service (SaaS), users are provided access to application software and databases. Cloud service providers manage the infrastructure and platforms that run the applications. SaaS is sometimes also called as "on-demand service of software" and is usually available on a pay-per-use basis. Cloud service providers install and operate application software in the cloud and cloud users can access the software from cloud clients. Cloud users need not manage the cloud infrastructure and platform where application runs. This reduces the need to install and run the applications on the cloud user's own computers. This simplifies maintenance and support.

Related Work

Data cloud is to share and analyze the data resources, storage resources and others in a wide network which is dynamic, heterogeneous and distributive. The data distributed across a grid must be available and accessible to several applications with a reasonable performance. It mainly focuses on analyzing massive data. In order to analyze the dynamic, real-time and online data, the whole data cloud system must be improved in order to enhance the access speed and reliability and safety and system's load balance. Therefore, how to choose the replication strategy for data cloud is particular important. Creating replica is to reduce access latency and bandwidth consumption, in other words, it is to reduce the average job execution time and improve the usage of cloud resources. These best data replication algorithms are based on some kind’s historical data access information and metadata. This information is in static condition. But, cloud is in dynamic environment. In cloud environment data replicate on a particular node. A data replica is best for some node at a certain point of time is not necessary to be best replication for another node at some different time. Because, workload, CPU capacity, changes in networks, etc. So, to select best data replication algorithm is still a problem. The main aim of this research is to provide a better technique for solving the drawbacks that currently exist in the literary works of data replica method in cloud environment. Here, we intend to propose Data Replication system based on data mining techniques. In proposed method, we will use a cloud computing. Combination of replication algorithm and job scheduling policy for data replication in the data cloud environment through monitoring all job process. The data replication will be done by identifying the frequently used data patterns in the large database of a node. This will be done by frequent pattern mining algorithm. The system availability and replica will be measured and identifying the location in which the replicated data will be stored. The replication will be stored based on the data failure probability and system availability. The popularity or frequency of the data will generate an adaptive threshold value for replication. The replicated data will be equally distributed among the nodes for easy access.

Securing the Cloud Data Protection and Access

Machine learning algorithms have gained lot of attention in field of computer science. Machine learning is the sub branch of artificial intelligence, that build up and studies the systems that can learn from data. For example, a machine learning system can be used trained on email messages that can learn to identify difference between spam and non-spam messages. After learning from known data, the model can then be used to classify new email messages into spam and non-spam folders. Machine learning uses prediction on the Three main requirements of database replication are the performance, the availability and the consistency of data. These requirements are in conflict with each other because a change for the benefit of one of the criterion implies a change (minimization) at the expense of the other criteria. The access to a replicated entity is
typically uniform with access to a single, non-replicated entity. The replication itself should be transparent to an external user. In addition, in a failure scenario, a failover of replicas is hidden as much as possible. The data replication in the cloud can be explained with the help of the architectural diagram. The architecture shows that it contains three major sections like User, Scheduling manager, and Replica manager. The users are normally the clients those who access the cloud from different locations. Each user can access the cloud independently and will provide different data access which has the property of replication. The particular task of each user is first given to the scheduling manager divides the task to the corresponding data centers through the replica manager based on the number of user using the particular data center to access the file without collision. The dynamic data replication strategy consists of The first stage is to identify which data file should be replicated and when to replicate in cloud computing to reduce the waiting time. We modify the popularity degree in the first stage of the D2RS algorithm. The modification of the popularity degree is based on double exponential moving average function. Cloud computing provides massive clusters for efficient large scale computation and data analysis. MapReduce [5] is a well-known programming model which was designed for improving the performance of large batch jobs on cloud computing systems. However, there is growing interest in employing MapReduce and its open-source implementation, called Hadoop, for various types of jobs. This leads to sharing a single Hadoop cluster between multiple users, which run a mix of long batch jobs and short interactive queries on a shared data set. Data mining, a relatively young and interdisciplinary field of computer science, is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The goal of data mining is to extract knowledge from a data set in a human understandable structure and involves database and data management, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post processing of found structure, visualization and online updating. The large set of evolving and distributed data can be handled efficiently by Incremental Data mining, Parallel Data mining, and Distributed Data Mining. The second stage of the D2RS algorithm is to find the required number of replicas and the third stage is to find where to place the new replica data files. A user's file system is different from every other user's of the cloud. Because the user knows his file system, he knows where his data is stored, but malicious insider is unaware of the user's file system and therefore searches for the data. Here we can profile [8] difference of search pattern from valid user to an intruder. Once we have finished indentifying malicious insider, system asks the user's of the cloud to enter security key (challenge question) which is known only to the valid user. Intruder obviously enters wrong key for known reasons. This is where we launch disinformation attack, which is assumed to be correct by the intruder, i.e. decoy information. [9] In this way we can stop unauthorized access of the data and minimize data theft in the cloud environment. All unsuccessful attempts are maintained in the log file for further analyses. We identify difference between normal files and decoy files by attaching HMAC to the user's file for maintain authenticity of the data. Thus SVM ensures to minimize data theft in the cloud when compared to other machine learning algorithms[7]. SVM has limitation on the size of the data sets, It cannot be used on large data sets.

EXTENDING SVM TO MULTI CLOUDS

Intercloud or multi cloud is interconnection of network of networks i.e. two or more number of clouds connected across the network. The Inter-cloud concept is based on the key factor that each single cloud has
limited physical resources. If a cloud saturates its computational and storage resources as part of infrastructure, or is requested to use resources which cannot be extended beyond its capacity, it will still be able satisfy such requests for service allocations sent from clients. The Inter-cloud theory would address situations, where each cloud will use computational, storage, or other resource of the infrastructures of other clouds. It is similar to mobile technology where operators implement roaming and interoperability to their mobile customers. Exchange of data in cloud, peering, and roaming introduce new business opportunities among cloud service providers. When there is a need for data transfer in inter-cloud’s environment, there has to be a provision for data security. Integrity, confidentiality, authentication should be maintained in multi clouds environment. [11] Cloud service providers should take adequate measures for reliable data transfer and data access. Using SVM in multi clouds environment helps to reduce data theft. In multi clouds large data sets [10] needs to be processed, we can use BIG DATA Analytics technique called HADOOP. HADOOP [12] is a product Apache software foundation. It is an open source framework which gained much popularity in recent times. Map and Reduce functions are the two most common techniques in HADOOP that can help in processing large data sets.

Conclusion

In this paper, we present a approach for securing personal and business data in the Cloud. We also propose ways to monitor data access patterns by profiling user behavior to determine abnormal data access. Once unauthorized data access or exposure is suspected and later verified, with security questions, we can inundate the malicious insider with decoy information and minimize data theft. Decoy documents are stored in the Cloud along with user's real data serve as sensors to detect illegitimate access. Preventive attacks that rely on disinformation technology can provide quite a good level of security in the Cloud.

References


Mr. Vellanki Vijayababu is a student of Sree Rama Institute of Technology & Science, Kuppenakunta, Penuballi, Khammam, AP, India. Presently he is Pursuing his M.Tech (CSE) from this college. His area of interests includes Information Security, Cloud Computing, Data Communication & Networks.

Mr. P.Nageswa Rao is an efficient teacher, received M.Tech from JNTU Hyderabad is working as an Assistant Professor in Department of C.S.E, Sree Rama Institute of Technology & Science, Kuppenakunta, Penuballi, Khammam, AP, India. He has published many papers in both National & International Journals. His area of Interest includes Data Communications & Networks, Database Management Systems, Computer Organization, C Programming and other advances in Computer Applications.

Mr.S.Suresh well known author and excellent teacher. He is working as a H.O.D for the Department of M.Tech., Computer Science & Engineering, Sree Rama Institute of Technology & Science, Kuppenakunta, Penuballi, Khammam. He has vast teaching experience in various engineering colleges. To his credit couple of publications both National& International conferences / journals. His area of Interest includes Data Warehouse and Data Mining, information security, Data Communications & Networks, Software Engineering and other advances in Computer Applications. He has guided many projects for Engineering Students.