MJRET Open Access
Multidisciplinary Journal of Research in Engineering and Tec
Journal homepage: www.mjret.in
ISSN:2348 - 6953

NEED OF PRIVACY PRESERVATION IN CLOUD DATABASES

[1]Niraja Jain, [2]Dr B Raghu, [3]Dr V Khanaa, [4]Dr A V Deshpande

[1]Research Scholar, Bharath University, Chennai
[2]Professor, SREC, Chennai, [3]Dean, Bharath University, Chennai
[4]Principal, SKNCOE, Pune

Abstract: *The data is growing enormously with every passing day. With the paradigm shift in the technology used to store, manage and retrieve the data from file storage to relational databases to today's cloud databases, the need to process the query efficiently while preserving the privacy had always been the priority. The dynamism of the query and the optimization of the resources used for query execution are the aspects that don't change in any situation. The timing constraint on the query execution could be instantaneous, urgent, leisurely or indefinite. But in almost all the scenarios the effective storage, manipulation and data retrieval is the key. While retrieving the data stored in cloud, privacy preservation is of utmost importance.*

We propose an experimental setup to test the effectiveness of some storage policies in cloud databases that may help in effective retrieval of the data while preserving the privacy. The paper discusses the principles of data privacy in the light of cloud computing, the Eucalyptus private cloud, the existing storage structure of Eucalyptus and the possible alterations to be made to it to achieve privacy preserved data retrieval.

Keywords: *Cloud databases, privacy preservation, data storage*

1. INTRODUCTION

Cloud computing technology had changed the way the data is stored and processed. The multidimensional growth in computing systems and technologies have resulted into advanced scalable, portable and large scale integrated systems and technologies. Data centers, virtualization, cloud and WEB2 technologies are frontiers of such growth. Computing has changed itself from being a product to a service that can be delivered to the consumers over internet through large scale data centers or cloud. the data centers used to create cloud services represents a significant investment in capital outlay and on-going cost. The private and public cloud platforms offer to deliver all the benefits of cloud computing technology to their customers. Databases had always been the critical part of this en-devour. E.F. Codd in 1970 [2] proposed the rules for relational databases resulting into efficient and optimized query processing. To achieve such effectiveness today's cloud databases need to

be made cloud computing compatible. Recent technologies are putting more demand on query optimization and data security with increasing scalability.

The clouds provide on demand resources or services over the internet, usually at the scale and with the reliability of a data center. The types of clouds can be categorized into architectural model, computing model, management model and payment model. Following are the major features of the systems implementing cloud with datacenter, virtualization and WEB2 interfaces [3].

1. Efficiency
2. Fault tolerance
3. Ability to run in heterogeneous environment
4. Ability to operate on encrypted data
5. Ability to interface with business intelligence products

1.1 Privacy issues in cloud:

It is believed that information security policies cover the data privacy as well. Of course the two terms are interrelated. But in context of cloud computing, data privacy has its own concerns. The concept of privacy varies among countries, cultures, and jurisdiction as well as on the infrastructure of the organization. Privacy rights are related to the collection, use, disclosure, storage and destruction of personal data [9]. Privacy preservation should be managed as part of the data used by the organization.

1.2 Impact of cloud on data privacy:

1. Generation of information
2. Use
3. Transfer
4. Transformation
5. Storage
6. Archival
7. Destruction

Fig.1: Impact of cloud on data privacy

1.3 Key privacy concerns in cloud:

- Organization's ability to provide the individual with access to all personal information and to comply with stated requests.
- How the existing privacy compliance requirements are impacted by the move to the cloud?
- Data transfer in cloud usually happens without the knowledge of organization resulting in potential violation of local law.
- How long the personal information will be retained and who enforces this policy in cloud?
- Does the destruction of data policy destroys data or just makes it inaccessible to the organization?

2. CHALLENGES IN EXISTING SYSTEM

The existing Walrus storage provides the list of private data to the users. The request to access public or protected data can be satisfied on demand of the client. But client need to specify the bucket name or the access URL for the bucket in advance [1]. Usually the time required to locate the private bucket is less than the time required to access public or protected bucket because access control list of the bucket needs to be searched to check whether or not the user can access it. And revoking the access policy to reduce the search time for the buckets may compromise the privacy of the buckets.

In a cloud scenario neither the data owner nor the cloud server can enforce the owner's access control policy. The reasons being maintaining the confidentiality and performance. The data owner instead needs to mediate every access request to filter the query result. This in turn nullifies the advantage of storing data at an external server. Therefore it is necessary to design a mechanism such that the data themselves enforce the restrictions on the set of users who can access it.

Both the privacy of the users accessing cloud services and of the data stored at cloud servers may be at risk since access requests could be exploited either by the cloud server or by a malicious observer to possibly infer the sensitive content of the accessed data. The query evaluation process is also at risk since the cloud server is not trusted and therefore can compromise the integrity of query results. Query results satisfy the integrity checks if they are correct (i.e. computed on genuine data), complete (i.e. computed over the whole data collection), and fresh (i.e. computed on the most recent version of the data). Correctness can be achieved by digital signatures. Defining authenticated data structures on the data may ensure completeness. However these data structures may be less flexible as they provide integrity guarantee only for queries operating on the attribute on which the structure has been defined. Freshness can be provided by making authenticated data structures dependent on a variable that changes over time.

3. METHODOLOGY

Imagine that a client has developed a web based search service that is available to the world for use through WEB2. Cloud computing will enable the developer to host this service remotely and can deal with the scale variability efficiently. As the business grows or shrinks, developer can acquire or release the resources easily and relatively inexpensively. On the other hand, implementation and maintenance of the data services that are scalable and adaptable to such dynamic conditions becomes a challenge. Especially when the data services are the compositions of the other possibly third party services (e.g., Google search or Yahoo Image search), these services becomes the data processing graphs that use the third party services as building blocks and invoke them during their execution. Running these data services under different QoS (Quality of Service) constraints as per the client's requirements further makes the system complex [7].

To decide upon which third party services to be used in processing correctly, making the optimal use of the resources available, satisfy all the QoS constraints is quite difficult [5]. To make data services scalable and adaptable to the cloud environment, the data flow needs to be optimized automatically. This is analogues to query optimization and execution in traditional databases. The basic building blocks of any query optimization algorithm are:

1. Query Evaluation Plan generation
2. Search strategy
3. Cost function.

Cloud services make easier for users to access their personal information from databases and make it available to services distributed across internet. This makes it vulnerable to security and privacy attacks [10]. Whenever a client uses a new cloud data service, it needs to establish the identity usually by filling up the online form and providing sensitive personal information. This leaves a trail of personal information that need to be properly protected else may be misused. This underlines the need to develop the methods for effectively querying and deriving insight from ensuing sea of heterogeneous data. A specific problem is to answer the keyword queries over large collections of heterogeneous data sources. Developing index structures to support querying hybrid data is difficult. With emerging new technologies having potential to create new data arrangement scenarios in which users join ad-hoc communities to create, collaborate, curate and discuss data online. Data virtualization is the presentation of data as an abstract layer, independent of underlying database systems, structures and storage. Database virtualization is the decoupling of the database layer which lies between the storage and application layers within the application stack.

Herein we propose to use the Eucalyptus private cloud architecture as the experimental setup for testing the effectiveness of privacy preservation strategy in storage clouds. Storage clouds allow the client to upload the data to the servers. The data is available as and when required with the implemented security policies and reliability. The most difficult task for the storage clouds is to identify the authorized user for the requested data from the huge collection. Cloud providers allow for direct access to the private data. It is expected that only the authorized clients from the list provided to get access to the private data be allowed and

others be denied the service by the data provider. We study the Eucalyptus private cloud architecture in the following section.

4. EUCALYPTUS OPEN-SOURCE PRIVATE CLOUD

Eucalyptus is a Linux-based open-source software architecture that implements efficiency-enhancing private and hybrid clouds within an enterprise's existing IT infrastructure [12]. Eucalyptus is an acronym for "Elastic Utility Computing Architecture for Linking Your Programs to Useful Systems." A Eucalyptus private cloud is deployed across an enterprise's "on premise" data center infrastructure and is accessed by users over enterprise intranet. Thus, sensitive data remains entirely secure from external intrusion behind the enterprise firewall.

4.1 Eucalyptus Components:

Each Eucalyptus service component exposes a well-defined language agnostic API in the form of a WSDL document containing both the operations that the service can perform and the input/output data structures. Inter-service authentication is handled via standard WS-Security mechanisms. There are five high-level components, each with its own Web-service interface, that comprise a Eucalyptus installation.

4.1.1 Cloud controller:

Cloud Controller (CLC) is the entry-point into the cloud for administrators, developers, project managers, and end-users. The CLC is responsible for querying the node managers for information about resources, making high level scheduling decisions, and implementing them by making requests to cluster controllers.

Functions:

1. Monitor the availability of resources on various components of the cloud infrastructure, including hypervisor nodes that are used to actually provision the instances and the cluster controllers that manage the hypervisor nodes.

2. Resource arbitration – deciding which clusters will be used for provisioning the instances.

3. Monitoring the running instances.

In short, CLC has a comprehensive knowledge of the availability and usage of resources in the cloud and the state of the cloud.

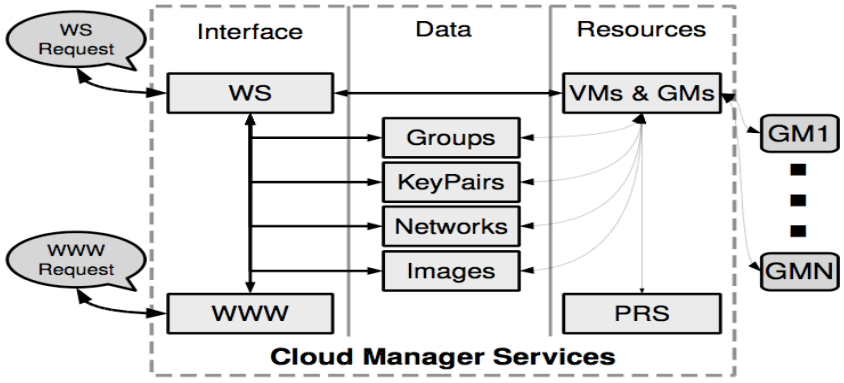

Fig.2: Overview of Cloud Controller services. Dark lines indicate the flow of user requests while light lines correspond to inter-service system messages.

4.1.2 Cluster Controller:

Cluster Controller (CC) generally executes on a cluster front-end machine or any machine that has network connectivity to both the nodes running NCs and to the machine running the CLC. CCs gather information about a set of VMs and schedules VM execution on specific NCs. The CC also manages the virtual instance network and participates in the enforcement of SLAs as directed by the CLC. All nodes served by a single CC must be in the same broadcast domain (Ethernet).

Functions:

1.To receives requests from CLC to deploy instances.
2. To decide which NCs to use for deploying the instances on.
3. To control the virtual network available to the instances.
4. To collect information about the NCs registered with it and report it to the CLC.

4.1.3 Node controller:

Node Controller (NC) is executed on every node that is designated for hosting VM instances. A UEC node is a VT-enabled server capable of running KVM as the hypervisor. UEC automatically installs KVM when the user chooses to install the UEC node. The VMs running on the hypervisor and controlled by UEC are called instances. The NC runs on each node and controls the life cycle of instances running on the node. The NC interacts with the OS and the hypervisor running on the node on one side and the CC on the other side. NC queries the operating system running on the node to discover the node's physical resources – the number of cores, the size of memory, and the available disk space. It also learns about the state of VM instances running on the node and propagates this data up to the CC.

Functions:

1. Collection of data related to the resource availability and utilization on the node and reporting the data to CC.
2. Instance life cycle management.

4.1.4 Storage controller:

Storage Controller (SC) implements block-accessed network storage (e.g., Amazon Elastic Block Storage -- EBS) and is capable of interfacing with various storage systems (NFS, iSCSI, etc.). An elastic block store is a Linux block device that can be attached to a virtual machine but sends disk traffic across the locally attached network to a remote storage location. An EBS volume cannot be shared across instances but does allow a snapshot to be created and stored in a central storage system such as Walrus, the Eucalyptus storage service.

Functions:

1. Creation of persistent EBS devices.
2. Providing the block storage over AoE or iSCSI protocol to the instances.
3. Allowing creation of snapshots of volumes.

4.2 Walrus:

EUCALYPTUS includes Walrus, a S3 compatible storage management service for storing and accessing user data as well as images. Walrus (put/get storage) allows users to store persistent data, organized as eventually-consistent buckets and objects. It allows users to create, delete, list buckets, put, get, and delete objects, and set access control policies. Walrus is interface compatible with Amazon's S3, and supports the Amazon Machine Image (AMI) image-management interface, thus providing a mechanism for storing and accessing both the virtual machine images and user data. Using Walrus, users can store persistent data, which is organized as buckets and objects. WS3 is a file-level storage system, as compared to the block-level storage system of Storage Controller.

5. CONCLUSION

For wide spread exploitation of the cloud technology, the effective and efficient solutions for protecting the privacy of the data stored in cloud infrastructure is very important. The study illustrate the fact that the EUCALYPTUS system has filled an important niche in the cloud-computing design space by providing a system that is easy to deploy atop existing resources, that lends itself to experimentation by being modular and open- source. In the existing system, the data access policies are not effective for public or protected data retrieval. Data confidentiality has to be balanced with query processing functions and performance. Most cloud platforms fail to support SQL queries as they are not designed to support structured data management. To improve query efficiency in cloud with preserving the privacy can be achieved by introducing dataflow optimization techniques and defining data descriptor based algorithms.

REFERENCES

[1] Amar More, Sarang Joshi , "Privacy preserving algorithm using effective data lookup organization for storage clouds" , IJCCSA August 2012, DOI: 10.5121/ijccsa.2012.2404

[2] E. F. Codd, "Code optimization rules and relational database", Communications of the ACM Volume 13 , Issue 6 (June 1970) Pages: 377 – 387, ISSN:0001-0782

[3] Daniel J. Abadi, "Data Management in the Cloud: Limitations and Opportunities", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2009

[4] Daniel Nurmi, Rich Wolski et al., " The Eucalyptus Open-source Cloud Computing System",

[5] Hitoshi Motsumoto, Yutaka Ezaki, "Dynamic resource management in cloud environment", FUJITSU Sci & Tech J, Vol 47, No 3, pg 270-276, July2011.

[6] Johann Christoph Freytag, "The basic principles of query optimization in relational database management system", ECRC, March 1989

[7] R. Agrawal et al., " The claremont report on database research", ACM SIGMOD Record, 37(3):9–19, 2008

[8] Rebecca Herald, "What is the difference between security and privacy?", CSI July 2002 Alert Newsletter

[9] Sabrina De, Sara Faresti et.al. , "Managing and accessing data in the cloud: Privacy risks and approaches", IEEE 2012.2404

[10] Shiyuan Wang et al., "Comprehensive framework for secure query processing on relational data in cloud", VLDB workshop on secure data management Nov 2010, doc 2010-25, University of California, Santa Barbara

[11] Tim Mather, Subra Kumaraswamy, and Shahed Latif, "Cloud security and privacy", O'reilly, Sept 2009

[12] Yohan Wadia, "The Eucalyptus Open-source Private Cloud"

[13] http://open.eucalyptus.com
[14] http://mirror.transact.net.au/pub/sourceforge/d/project/de/deduplication/papers/EucalyptusUserGuide.v1.final.03.23.pdf
[15] http://www.eucalyptus.com/resources/cloud-myths-dispelled
[16] http://www.cca08.org/papers/Paper32-Daniel-Nurmi.pdf