BACHELOR OF COMPUTER APPLICATIONS

SEMESTER 6

DCA3243

CLOUD COMPUTING

Unit 7

Cloud Data Management

Table of Contents

1. INTRODUCTION

By now, you must be familiar with the concept of cloud computing and the various technologies of cloud computing. As we all are aware, cloud computing alters the IT landscape; data is the only thing that organisations maintain control over. IT people tend to think that their job is to manage technology and systems, yet data (not technology) is something that enterprises must manage as cloud computing becomes dominant.

A typical representation of the cloud is an on-demand, reliable service provided over the Internet with easy access to virtually infinite computing, storage and networking resources. Through very simple web interfaces and at small incremental costs, users can outsource complex tasks, such as data storage application deployment, to very large data centres operated by cloud providers. Thus, the complexity of managing the software/hardware infrastructure gets shifted from the users' organisation to the cloud provider.

For managing data, cloud providers could rely on relational DBMS technology, all of which have distributed or parallel versions. Since data management in the cloud is becoming a major research direction, cloud providers have developed new, proprietary technologies, typically with specific, simple applications. In this unit, we will discuss different clouds.

Security issues, data control mechanisms and cloud data storage as a service.

1.1 Objectives:

After reading this unit, you should be able to:

- *Explain security issues of cloud data.*
- *Analyse different types of data control techniques.*
- *Identify cloud databases and data stores.*
- *Describe data storage as a service.*

2. DATA SECURITY

As we all know, data security is the practice of keeping data protected from fraud and unauthorised access. The focus behind data security is to ensure privacy while protecting personal or corporate data. Under cloud computing, **security** (sometimes referred to simply as "cloud security") is an evolving sub-domain of computer security, network security, and, more broadly, information security. Per Cloud Security Alliance's definition of cloud computing with that of the National Institute of Standards and Technology's (NIST) definition, includes on-demand self-service, broadband network access, resource pooling, rapid provisioning and scalability, and metered usage. The massive, distributed Internet architecture of cloud computing has been influenced to provide data redundancy, faster access times, and rapidly scalable service to support the high demand for next-generation web technology. As a result, security has turned into the major concern of cloud computing.

Security issues related to cloud computing fall into two broad categories:

- Security issues faced by cloud providers and
- security issues faced by their customers.

NIST divides security issues faced by cloud providers into three categories:

- **Software-as-a-service** (SaaS): (applications supplied by the service provider);
- **Platform-as-a-service** (PaaS): (tools and programming languages supported by the provider for customers to deploy their own applications);
- **Infrastructure-as-a-service** (provider supplies hardware platforms within its network for customer use).

In most cases, the provider must ensure that their infrastructure is secure and that their client's data and applications are protected, while the customer must ensure that the provider has taken the proper security measures to protect their information.

Cloud computing technology itself is a very secure one. However, companies must carefully consider the suggestions of massively scalable design, storage, and computing. This is especially true if those services are outsourced to cloud providers and not directly under company control. Organisations might consider not only the legal implications of computing across national boundaries but also those of an international customer base, as well as the

challenges of ensuring that all copies of data throughout the cloud infrastructure uniformly fall under the appropriate policies. Most of the challenges obviously exist with public cloud computing and open cloud computing, where companies use resources outside their control from one or many providers to provide transparently scalable computing resources. However, private clouds do pose risks to enterprises that are not prepared for the complexities of this infrastructure. As resources are consumed for cloud computing tasks and then returned to the cloud for the next task, there is a risk that data permanence or audit trails are not sufficient. Organisations must plan carefully when constructing cloud computing environments to ensure that the flexibility and scalability do not outshine the necessity for risk-tolerant implementation. Securing your data in the cloud is done by implementing any of the following:

- Access control lists to define the permissions attached to the data objects
- Storage encryption to protect against unauthorised access at the data centre
- Transport level encryption to protect data when it is transmitted
- Firewalls to include web application firewalls to protect against outside attacks launched against the data centre
- Hardening of the servers to protect against known and unknown vulnerabilities in the operating system and software
- Physical security to protect against illegal physical access to data

The above implementation can be classified into three key components, which are related to data security and privacy.

- Data location
- Data control
- Securing data for transport

Now, we will discuss each component separately.

3. DATA LOCATION

When you use the cloud, you probably won't know exactly where your data is hosted. In fact, you might not even know which country it will be stored in. Evidently, data relocations allow you to specify the target storage and path. However, in the case of the cloud, unless you ask for it, you have no control over where your data will reside. Hence, while selecting a cloud vendor, it is important to check whether they allow and facilitate this choice. Customers might have compliance reasons to insist on data location within a region, and not having a strategy in place could mean loss of business.

As you are responsible for private and valuable data of your organisation, personal or company information, you need to know where data will be located at all times. In the cloud environment, location matters, especially from a legal standpoint. In the legal world, location is most frequently associated with jurisdiction. The concept of "jurisdiction" is allied with the power of a judge or government entity to declare authority over the persons or things involved in an action and to make a decision about a specific issue or set of facts. Jurisdiction is not necessarily exclusive. Several countries or courts may have concurrent jurisdiction over a matter. Indeed, complainants frequently argue about who has jurisdiction over their dispute. In the cloud environment, where a piece of equipment is located may have significant consequences on the ability of a court or other government authority to assert jurisdiction over that piece of equipment and, in the case of a server, over the data contained in that server.

If the cloud that hosts your data has servers in a foreign country, the laws of that foreign country may govern your data when stored on that server. As a result, many important foreign laws may govern your data. For example, the European Union places strict limits on what data can be stored on its citizens and for how long. Many banking regulators also require customers' financial data to stay in their home country. Many compliance regulations require that data not be intermixed with other data, such as on shared servers or databases. Consider the following cloud computing legal issues that stem from data location. So, when you are going to contact your cloud service vendor, ask them where your data will be stored or processed. Based on this, the provider should make a contractual commitment to obey local privacy requirements on behalf of their customers.

SELF-ASSESSMENT QUESTIONS – 1

1. SaaS stands for.
2. Cloud computing technology itself is a very secure one. (True/ False)
3. The concept of "_____" is allied with the power of a judge or government entity to declare authority over the persons or things involved in an action and to decide about a specific issue or set of facts.

4. DATA CONTROL

As we have seen, security is a major concern of cloud computing, so to make sure that your data in the cloud can be trusted by government policies is called data control in the cloud. So, cloud data are subject to being held 'hostage' by the cloud provider.

For example, suppose you are working with a word processor using your cloud service and want to store your document in the cloud. These documents belong to your company, and you expect to control access to those documents. Without your consent, no one can access it. Due to a software bug, privacy has been violated, resulting in a malfunctioning access control. This is completely undesirable. So, to handle such a situation, you must comprehend your' cloud provider's level of controls and how these controls can be audited.

Now, we will discuss different types of controls aimed at ensuring the completeness and accuracy of data input, output and processing. The different types of data controls are given below:

- **Access controls:** To guarantee that only those who are authorised to access the data can do so. Sensitive data must also be protected in storage and transfer. Encrypting the data can help to do this.
- **Input validation controls:** To confirm that all data input to any system or application is whole, correct, and realistic.
- **Output reconciliation controls:** To ensure that data can be reconciled from input to output.
- **File controls:** To make certain that data are operated accurately in any type of file (structured and unstructured).

- **Change management controls:** To ensure that data can't be changed without proper agreement.
- **Processing controls:** To safeguard that data is processed completely and accurately in an application.
- **Data destruction controls:** To ensure that when data is permanently deleted, it is deleted from everywhere — including all backup and redundant storage sites.
- **Backup and recovery controls:** Many security openings come from difficulties in data backup. Maintaining physical and logical controls over data backup is very crucial.

5. SECURING DATA FOR TRANSPORT

In today's technology environment, encryption of data within the enterprise has become a minimal requirement. However, more recently, organisations that have begun outsourcing various IT resources to the cloud have begun to discover that the recognised techniques for securing data that they have relied upon for so long are inadequate in a cloud environment. In spite of this, there are two main faces to protecting that data, whether an organisation stores data locally or in the cloud, i.e., data must be confined when it is relaxed and when it is in action.

Protecting data in action refers to upholding data that is being sent across the wire. This is completely different from storage encryption. When a user attempts to access data from an encrypted volume over a network, there are processes that occur to ensure that the user has the necessary rights. The effective way to protect your data is to provide storage encryption and transport encryption. Storage encryption is not a new one, and transport encryption is an addition to that. However,, cloud service providers provide the least protection for stored data where they provide different transport level encryption like HTTPS, TLS, IPSec, etc., for protecting data in motion.

Considering a situation where data in the cloud moving from one point A to another point B may take on three different forms,

- Between two clouds
- Within a cloud environment

- Over the public Internet between an enterprise and a cloud provider. Using any method, data encryption can be done, and isolation of your data from other companies' data can be done by different security processes.

As we have already discussed in the previous unit, a VPN (Virtual Private Network) can manage data security in a cloud environment when it is transmitting. A VPN basically makes the public network your own private network, which is an alternative to using dedicated connectivity. A well-designed VPN needs to incorporate two things:

- A **firewall** to act as an obstruction between the public Internet and any private network.
- **Encryption** to protect your sensitive data from hackers; only the computer that you send it should have the key to decode the data.

Your level of concern about security may vary depending on the control requirements for your data. It is not like the cloud providers are not aware of the protection of data. Still, the degree of protection that is provided depends on what type of cloud is being used (SaaS, IaaS, PaaS, etc.) and which company is providing the service.

SELF-ASSESSMENT QUESTIONS – 2

4. _____ Confirms all data input to any system or application are whole, correct, and realistic to ensure the completeness and accuracy of data input-output and processing.
5. A VPN basically makes the public network your own private network as an alternative to using dedicated connectivity. (True/False)
6. A _____ to act as an obstruction to between the public Internet and any private network.

6. SCALABILITY AND CLOUD SERVICES

By now, you must be familiar with the security and privacy of cloud data. Now, we will discuss the reasons behind the demand for cloud services. The economics of cloud computing are compelling, and increasing numbers of application developers and application owners are accepting the model. So, the most challenging factor with the cloud is how to manage and scale your data across a distributed infrastructure.

For example, earlier, almost all public video was stored by TV networks. No one could be able to access those data based on their requirement. But nowadays, a volatile number of videos are currently available through YouTube, which was unimaginable prior to its creation in 1995. Today, you can store videos, watch videos, and search for videos by using YouTube as your video provider (to handle the streaming of the video to your Web site) anytime.

Let us go through the up-and-coming technologies for organising these increasing volumes and variety of data:

- Resources to support large-scale processing in the cloud
- Databases and data stores in the cloud
- Data archiving in the cloud

6.1 Large-scale processing

As data collection and storage technologies progress, large data sources have become omnipresent. Today, organisations regularly collect terabytes of data on a daily basis with the intent of gathering non-trivial insights into their business processes. To benefit from these advances, it is very important that data mining and machine learning techniques scale to such proportions. The attraction of cloud computing is its flexibility. You can add as much capacity as you need to process and analyse your data. The data might be processed on clusters of computers. This means that the analysis is occurring across machines. Companies are considering this approach to help them manage their supply chains and inventory control or consider the case of a company processing product data from across the country to determine when to change a price or introduce a promotion. The data might come from the point-of-sale (POS) systems across multiple stores in multiple states. POS systems

generate a lot of data, and the company might need to add computing capacity to meet demand.

This model is large-scale, distributed computing, and a number of frameworks are emerging to support this model, including:

- **MapReduce** is a patented software framework introduced by Google to support distributed computing on large sets of data. It is designed to take advantage of cloud resources. This computing is done across large numbers of computers, called clusters or as a grid. Each cluster is referred to as a node. MapReduce can deal with both structured (database) and unstructured (file system) data. Users specify a map function that processes a key/value pair to generate a set of intermediate pairs and a reduction function that merges these pairs.
- **Apache Hadoop** is a data-intensive open-source distributed computing platform written in Java inspired by Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed on any node in the cluster. The cloud-based Hadoop system and associated Apache systems have become very popular for processing large-scale data for data mining and related applications. It creates a computer pool, each with a Hadoop file system. It then uses a hash algorithm to cluster data elements that are similar. Hadoop can create a map function of organised key/value pairs that can be output to a table, to memory, or to a temporary file to be analysed. Three copies of the data exist so that nothing gets lost.

6.2 Databases and Data Stores

In cloud computing, handling a huge amount of data is a big challenge, so it is not unexpected that new database technologies are being developed to support this kind of computing. People seem unwilling to run relational databases in a cloud instance. Probably a part of this is the license cost; replicating existing data centre commercial databases in cloud instances can get expensive. That's at least one reason why the Amazon Relational Database Service (RDS) uses MySQL. While not free for commercial use, MySQL support costs substantially less than that of traditional commercial databases. Promising non-relational platforms are compelling options for organisations looking to overcome the cost and scalability limitations of relational database systems. Another cause may be performance. This can be slow when you're executing complex queries that involve a join across a distributed environment.

Additionally, in an old-style database cluster, data must either be replicated across the boxes in the cluster or partitioned between them. According to other database experts, this makes it hard to provision servers on demand.

Most IT groups will probably prefer that data be persisted directly to the enterprise relational databases in the local data centre, but the latency is such that it would prohibit the scalability of enterprise applications. To overcome these limitations, some large cloud providers accepted the nascent no-SQL movement and developed their own databases. Some of them are listed below.

Cloud-based SQL: Microsoft has introduced a cloud-based SQL relational database called SQL Database (SDS). SDS provides data storage by using a relational model in the cloud and access to that data from the cloud and client applications. It runs on the Microsoft Azure services platform. The Azure platform is an Internet-scale cloud-services platform hosted in Microsoft data centres; the platform provides an operating system and a set of developer services. In a report on SQL Azure, analyst firm Forrester has quoted: "Most customers stated that SQL Azure delivers a reliable cloud database platform to support various small to moderately sized applications as well as other data management requirements such as backup, disaster recovery, testing, and collaboration."

Google Bigtable: Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. Many projects at the Google store data in Bigtable, including web indexing, Google Earth, and Google Finance. This hybrid is sort of like one big table. Because tables can be large, they're split at row boundaries into tablets, which might be 100 megabytes or so. MapReduce is often used to generate and modify data stored in Bigtable. Bigtable is also the data storage vehicle behind Google's App Engine (a platform for developing applications).

Amazon SimpleDB: Amazon SimpleDB is a highly available, flexible, and scalable non-relational data store that offloads the work of database administration. This Web service is for indexing and querying data. It is used with two other Amazon products to store, process, and query data sets in the cloud. Amazon relates the database to a spreadsheet in that it has columns and rows with attributes and items stored in each. Unlike a spreadsheet, each cell

can have multiple values, and each item can have its own set of associated attributes. Amazon then automatically indexes the data.

As organisations expand their Web and cloud deployments, lots of open-source databases are being developed every day. Some of them have been included below.

- Mango DB v.1.8.0: high performance document-centric database released by 10Gen.
- Drizzle 7: fork of MySQL database developed by Drizzle maintainers.
- CouchDB: is a document-oriented database developed by Apache.
- LucidDB: is the first and only open-source RDBMS purpose-built entirely for data warehousing and business intelligence, developed by the Eigenbase Project.

6.3 Data Archival

We all are aware of the definition of data archiving, which is nothing but the process of moving data that is no longer actively used to a separate data storage device for long-term retention; companies of all sizes are rethinking their approach to data retention, where data volumes continue to grow exponentially. Companies have to hold onto data for long periods of time. Archiving data to value tiers of storage, such as storage clouds, is now being viewed as a desirable alternative to traditional storage systems. The cloud has different data archiving models. In some models, the archive may be available on demand. In others, this may not be the case. For example,

The ABC Data Archive Cloud Store Option extends the capabilities of ABC Data Archive to optimally store and easily and efficiently access inactive data from databases and enterprise applications in the cloud.

Key features of the ABC Data Archive Cloud Store Option include:

- Efficient, cost-effective cloud archiving
- Optimized access to data archived in the cloud.
- A platform-neutral approach to managing data archived in the cloud.
- Enterprise-grade disaster recovery and high availability
- Centralized management of on-premises and cloud archiving through a single-user interface

7. STORAGE AS A SERVICE

We have discussed security, location, and control of cloud data. Now, we will discuss cloud storage. We have seen in the previous section that the security of stored data and data in movement is a major concern when storing sensitive data at a cloud storage provider. Enterprises are looking for reasonable alternatives to address the continuous growth of storage while addressing disaster recovery and compliance initiatives. At the same time, users require easy and immediate provisioning of storage as they need it and access to their information how and when they need it. Cloud storage addresses these needs by reducing the cost of storing files, automating business processes, and providing omnipresent file access to employees globally. So, networked online storage, where we can store data on multiple virtual servers, is called cloud storage. This storage is hosted by a third party rather than being hosted on dedicated servers. Suppose an organisation requires some storage from a data centre for their storage needs. Now, the data centres virtualise the resources according to the requirements of the customer and represent them as storage pools. Organisations can access their cloud storage services through a web service application programming interface or through a web-based interface. Still, physically, the resource may span across multiple servers. So, they can access their storage as a service from a provider.

There is no shortage of popular cloud storage services. Amazon announced cloud storage for music. Google is reportedly working on a massive music streaming service that will reside in the cloud. New reports indicate that Apple is working on its own cloud music service that might even beat Google in the market. Some of the online storage services have been

Discussed below, which has a good track record of keeping your data safe while providing you easy access to your files from wherever you are.

Live Mesh: The online storage component of Live Mesh is only one part of Microsoft's latest venture into cloud computing, but it is also one of its most compelling features at this point. Live Mesh gives you 5GB of online storage and an online desktop that looks a lot like Windows Vista. You can upload any type of file to Live Mesh, but you cannot edit any of your files through the online desktop. In the future, though, we expect Microsoft to start adding more of these features. One of the main reasons for its popularity is that it constantly watches for changes in the folders you are synchronising to and updates them automatically. In

addition, you can share folders with friends, allowing you to collaborate on projects. Live Mesh works on both Windows PCs and Macs.

Zumo: This is a cloud-based file synchronisation and storage service. The service enables users to store and sync files online and between computers using their Hybrid Cloud storage solution. Zumo Drive has a cross-platform client between (Windows, Mac, Linux, iOS, Android, and Palm web OS)

Wuala: This is a free, secure online storage that allows its users to store, backup, and access files from anywhere and to share files easily with friends, groups, and the world. Tens of thousands of users and thousands of communities around the world are already actively sharing millions of files, growing quickly.

JungleDisk: This is available for free - and technically, it is not even an online storage service. Instead, it provides a front end to Amazon's S3 storage service. JungleDisk also lets you map your Amazon S3 storage space as a network drive on your computer so that you can just drag and drop files back and forth between your online storage and your local desktop. JungleDisk is available for Windows, Mac OSX, and Linux.

SELF-ASSESSMENT QUESTIONS – 3

7. _____ is a patented software framework introduced by Google to support distributed computing on large sets of data.
8. _____ is the first and only open-source RDBMS purpose-built entirely for data warehousing and business intelligence, developed by the Eigenbase Project.
9. Microsoft has introduced a cloud-based SQL relational database called Bigtable. (True/False)
10. Cloud storage is hosted by a third party rather than being hosted on dedicated servers. (True/False)
11. Live Mesh gives us _____ online storage and an online desktop that looks a lot like Windows Vista.
    a) 2 GB
    b) 4 GB
    c) 5 GB
    d) 8GB
12. Organization can access their cloud storage services through a _____.

8. SUMMARY

Let us now recapitulate the important points discussed in this unit.

- Cloud computing, associated with a number of securities matters, falls into two broad categories: Security issues faced by cloud providers and security issues faced by their customers. NIST divides security issues faced by cloud providers into three categories: SaaS, PaaS, and IaaS.
- In the cloud environment, location matters, especially from a legal standpoint. In the legal world, location is most frequently associated with jurisdiction.
- In the case of cloud data control, different types of controls aimed to ensure the completeness and accuracy of data input-output and processing, such as Input validation controls, Output reconciliation controls, File controls, Change management controls, Processing controls, Data destruction controls.
- A VPN basically makes the public network your own private network as an alternative to using dedicated connectivity. A well-designed VPN needs to incorporate two things: firewall encryption.
- In the case of distributed computing to handle large-scale cloud data, MapReduce and Apache Hadoop are used, where MapReduce software is used.is a patented framework, and Apache Hadoop is open-source.
- Most IT groups would probably prefer that data be persisted directly to the enterprise relational databases in the local data centre, but the latency is such that it would prohibit the scalability of enterprise applications. To overcome these limitations, some large cloud providers accept the nascent no-SQL movement and develop their own databases, E.g.Cloud-based SQL, GoogleBigtable, etc.
- Data archiving is the process of moving data that is no longer actively used to a separate data storage device for long-term retention.
- Cloud storage is networked online storage where we can store data on multiple virtual servers, and this is generally hosted by a third party. Some of the examples of cloud storage services are LiveMesh, JungleDisk, Zumo, Wuala, etc.

9. TERMINAL QUESTIONS

1. Explain different types of security issues faced by cloud providers.
2. What do you mean by 'jurisdiction' in the case of cloud data location?
3. Discuss different types of controls aimed at ensuring the completeness and accuracy of data input, output, and processing.
4. Define large-scale processing in the cloud.
5. List out a few cloud databases that are influenced by the no-SQL movement.
6. Explain data archival in the cloud.
7. Specify two open-source cloud storage services.

10. ANSWERS

Self-Assessment Questions

1. Software As a Service
2. True
3. Jurisdiction
4. Input validation controls
5. True
6. firewall
7. MapReduce
8. LucidDB
9. False
10. True
11. 5 GB
12. web service application programming interface

Terminal Questions

1. NIST divides security issues faced by cloud providers into three categories: Software-as-a-service, Platform-as-a-service, and Infrastructure-as-a-service. Refer to section 7.2 for more details.
2. Jurisdiction is allied with the power of a judge or government entity to declare authority over the persons or things involved in an action and to decide about a specific issue or set of facts. Refer to section 7.3 for more details.

3. There are different types of controls aimed at ensuring the completeness and accuracy of data input-output and processing, like access control. Refer to section 7.4 for more details.
4. Today, organisations regularly collect terabytes of data on a daily basis with the intent of gathering non-trivial insights into their business processes. Refer to section 7.6.1 for more details.
5. Xeround is a cloud database which is influenced by the no-SQL movement. Refer to section 7.6.2 for more details.
6. Data archiving is the process of moving data that is no longer actively used to a separate data storage device for long-term retention. Refer to section 7.6.3 for more details.
7. Wuala is an open-source cloud storage service. Refer to section 7.7 for more details.

11. REFERENCES

- Judith Hurwitz, Robin Bloor, Marcia Kaufman, Fern Halper.*Cloud Computing for Dummies*.Wiley Publishing, Inc.

E-References:

- http://www.informatica.com/products_services/app_info_life_manage/options/cloud_store_option/Pages/index.aspx
- http://www.francoisegilbert.com/2011/04/cloud-computing-legal-issues- data-location/
- http://www.readwriteweb.com/cloud/2011/01/7-cloud-based-database-service.php
- http://pvarhol.wordpress.com/2009/12/16/the-cloud-data-store-dilemma/
- http://en.wikipedia.org/wiki/Cloud_storage
- http://www.cloudtweaks.com/2010/08/10-very-popular-cloud-storage- service-must-haves-for-businesses/
- http://www.readwriteweb.com/archives/free_online_storage_services.php