University of Science and Technology
Faculty of Computer Science and Information Technology
Information Security Master Program
Course Title: Cloud Computing Security

Lecture (3): Data Security

Reference: Cloud Computing Security and Privacy, by Tim Mather, Subra and Lattif, -------Chapter (4)

Instructor: Prof. Noureldien Abdelrahman

This lecture describes several aspects of data security, including:
• Data-in-transit
• Data-at-rest
• Processing of data, including multitenancy
• Data lineage
• Data provenance/Source
• Data remanence

3.1 Aspects of Data Security

3.1.1 Data-in-transit:

To insure the security of data in transit it is not only important to use a vetted encryption algorithm, but also it is important to ensure that the service protocol provides confidentiality as well as integrity (e.g., FTP over SSL [FTPS], Hypertext Transfer Protocol Secure [HTTPS], and Secure Copy Program [SCP])—particularly if the protocol is used for transferring data across the Internet.
Encrypting data and using a non-secured protocol (e.g., FTP or HTTP) can provide confidentiality, but does not ensure the integrity of the data.

3.1.2 Data-at-rest:

Although using encryption to protect data-at-rest might seem obvious, the reality is not that simple. If you are using an IaaS cloud service (public or private) for simple storage (e.g., Amazon's Simple Storage Service or S3), encrypting data-at-rest is possible—and is strongly suggested.
However, encrypting data-at-rest that a PaaS or SaaS cloud-based application is using (e.g., Google Apps, Salesforce.com) is not always feasible. Data-at-rest used by a cloud-based application is generally not encrypted, because encryption would prevent indexing or searching of that data.
Generally speaking, with data-at-rest, **the economics of cloud computing** are such that **PaaS based applications and SaaS applications is to use a multitenancy architecture.**

In other words, data, when processed by a cloud-based application or stored for use by a cloud-based application, is commingled with other users' data (i.e., it is typically stored in a massive data store, such as Google's BigTable).

Although applications are often designed with features such as data tagging to prevent unauthorized access to commingled data, unauthorized access is still possible through exploit of an application vulnerability.

Although an organization's data-in-transit might be encrypted during transfer to and from a cloud provider, and its data-at-rest might be encrypted if using simple storage (i.e., if it is not associated with a specification application), an organization's data is definitely not encrypted if it is processed in the cloud (public or private). ***For any application to process data, that data must be unencrypted.***

In June 2009, IBM announced that one of its researchers, working with a graduate student from Stanford University, had developed a fully homomorphic encryption scheme which allows data to be processed *without being decrypted.*

3.1.3 Data Lineage

Whether the data of an organization has been put into the cloud is encrypted or not, it is useful and might be required (for audit or compliance purposes) to know exactly where and when the data was specifically located within the cloud.

For example, the data might have been transferred to a cloud provider, such as Amazon Web Services (AWS), on date *x*1 at time *y*1 and stored in a bucket on Amazon's S3 in *example1.s3.amazonaws.com*, then processed on date *x*2 at time *y*2 on an instance being used by an organization on Amazon's Elastic Compute Cloud (EC2) in *ec2-67-202-51-223.compute-1.amazonaws.com*, then restored in another bucket, *example2.s3.amazonaws.com*, before being brought back into the organization for storage in an internal data warehouse belonging to the marketing operations group on date *x*3 at time *y*3.

Following the path of data (mapping application data flows or data path visualization) is known as *data lineage*, and it is important for an auditor's assurance (internal, external, and regulatory). However, providing data lineage to auditors or management is time-consuming, even when the environment is completely under an organization's control. Trying to provide accurate reporting on data lineage for a public cloud service is really not possible.

3.1.4 Data Provenance

Even if data lineage can be established in a public cloud, for some customers there is an even more challenging requirement and problem: proving data provenance—not just proving the integrity of the data, but the more specific provenance of the data

Data provenance documents the inputs, entities, systems, and processes that influence **data** of interest, in effect providing a historical record of the **data** and **its origins**.

How do you prove data provenance in a cloud computing scenario when you are using shared resources? Those resources are not under your physical or even logical control, and you probably have no ability to track the systems used or their state at the times you used them—even if you know some identifying information about the systems (e.g., their IP addresses) and the "general" location (e.g., a country, and not even a specific data center).

3.1.5 Data Remanence

Data remanence is the residual representation of digital **data** that remains even after attempts have been made to remove or erase the **data**. This residue may be due to data being left intact by a nominal delete operation, or through physical properties of the storage medium.

Data remanence may make inadvertent disclosure of sensitive information possible. The risk posed by data remanence in cloud services is that an organization's data can be inadvertently exposed to an unauthorized party—regardless of which cloud service you are using (SaaS, PaaS, or IaaS). When using SaaS or PaaS, the risk is almost certainly unintentional or inadvertent exposure. Various techniques have been developed to counter **data remanence**. These techniques are classified as clearing, purging/sanitizing, or destruction.

In spite of the increased importance of data security, the attention that cloud service providers (CSPs) pay to data remanence is strikingly low. Many do not even mention data remanence in their services.

3.2 Data Security Mitigation

Although data-in-transit can and should be encrypted, any use of that data in the cloud, beyond simple storage, requires that it be decrypted. Therefore, it is almost certain that in the cloud, data will be unencrypted. And if you are using a PaaS-based application or SaaS, customer-unencrypted data will also almost certainly be hosted in a multitenancy environment (in public clouds). Add to that exposure the difficulties in determining the data's lineage, data provenance—where necessary—and even many providers' failure to adequately address such a basic security concern as data remanence, and the risks of data security for customers are significantly increased.

So, what should you do to mitigate these risks to data security? The only viable option for mitigation is to ensure that any sensitive or regulated data is not placed into a public cloud (or that you encrypt data placed into the cloud for simple storage only).

3.3 Provider Data and Its Security

In addition to the security of customers own data, customers should also be concerned about what data the provider collects and how the CSP protects that data. What metadata does the provider have about customer data, how is it secured, and what of it is accessible to the customer? As your volume of data with a particular provider increases, so does the value of that metadata.

Additionally, your provider **collects and must protect a huge amount** of security-related data. For example, at the network level, your provider should be collecting, monitoring, and protecting firewall, intrusion prevention system (IPS), security incident and event management (SIEM), and router flow data. At the host level your provider should be collecting system logfiles, and at the application level SaaS **providers should be collecting application log data, including authentication and authorization information**.

What data your CSP collects and how it monitors and protects that data is important to the provider for its own audit ***purposes Additionally, this information is important to both providers and customers in case it is needed for incident response and any*** digital forensics required for incident analysis.

3.4 Storage Security

For data stored in the cloud (i.e., storage-as-a-service), we are referring to IaaS and not data associated with an application running in the cloud on PaaS or SaaS. The same three information security concerns are associated with this data stored in the cloud (e.g., Amazon's S3) as with data stored elsewhere: confidentiality, integrity, and availability.

3.4.1 Confidentiality

When it comes to the confidentiality of data stored in a public cloud, you have two potential concerns. First, what access control exists to protect the data? Access control consists of both **authentication and authorization**.

The second potential concern is: How is the data that is stored in the cloud actually protected? For all practical purposes, protection of data stored in the cloud involves the use of encryption. So, is a customer's data actually encrypted when it is stored in the cloud? And if so, what encryption algorithm, and with what key strength? It depends, and specifically, it depends on which CSP you are using. For example, EMC's MozyEnterprise does encrypt a customer's data.

However, AWS S3 does *not* encrypt a customer's data. Customers are able to encrypt their own data themselves prior to uploading**, but S3 does not provide** encryption.

Another confidentiality consideration for encryption is key management. How are the encryption keys that are used going to be managed—and by whom? Are you going to manage your own keys? Hopefully, the answer is yes, and hopefully you have the expertise to manage your own keys.

3.4.2Integrity

In addition to the confidentiality of your data, you also need to worry about the integrity of your data. Confidentiality does not imply integrity; data can be encrypted for confidentiality purposes, and yet you might not have a way to verify the integrity of that data. Encryption alone is sufficient for confidentiality, but integrity also requires the use of message authentication codes (MACs).

The simplest way to use MACs on encrypted data is to use a block symmetric algorithm (as opposed to a streaming symmetric algorithm) in cipher block chaining (CBC) mode, and to include a one-way hash function.

Another aspect of data integrity is important, especially with bulk storage using IaaS. Once a customer has several gigabytes (or more) of its data up in the cloud for storage**,** how does the customer check on the integrity of the data stored there? **There are IaaS transfer costs associated with moving data into and back down from the cloud.**

What a customer really wants to do is to **validate the integrity of its data while that data remains in the cloud**—without having to download and reupload that data. This task is even more difficult because it must be done in the cloud without explicit knowledge of the whole data set. Customers generally do not know on which physical machines their data is stored, or where those systems are located. Additionally, that data set is probably dynamic and changing frequently. Those frequent changes obviate the effectiveness of traditional integrity insurance techniques.

3.4.3 Availability

Assuming that a customer's data has maintained its confidentiality and integrity, you must also be concerned about the availability of your data. There are currently three major threats in this regard—none of which are new to computing, but all of which take on increased importance in cloud computing because of increased risk.

The first threat to availability is network-based attacks. The second threat to availability is the CSP's own availability. No CSPs offer the sought-after "five 9s" (i.e., 99.999%) of uptime. A

customer would be lucky to get “three 9s” of uptime. The **third risk is absent of data backup a**s cloud storage does not mean the stored data is actually backed up. Some cloud storage providers do back up customer data, in addition to providing storage. However, many cloud storage providers do not back up customer data, or do so only as an additional service for an additional cost.

All three of these considerations (confidentiality, integrity, and availability) should be encapsulated in a CSP’s service-level agreement (SLA) to its customers