International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Data Lineage in Malicious Environment

Ashish Satish Tayde

Department of Information Technology, B. K. Birla College Of Arts, Science & Commerce (Autonomous), Kalyan, India.

ABSTRACT:

Leakage of data is one of the most serious security threats that organizations face in this era. This thread is also outspreading in our personal lives. Personal information which we save in our smartphones is not secure many applications have access to our data which includes our media, contacts and many other things and is accessible to many third-party applications.

Introduction:

The digital era, information leakage through unintentional exposures, or intentional sabotage by disgruntled employees and malicious external entities, present one of the most serious threats to organizations.

According to an interesting chronology of data breaches maintained by the Privacy Rights Clearinghouse (PRC), in the United States alone, records have been breached from data breaches made public since It is not hard to believe that this is just the tip of the iceberg, as most cases of information leakage go unreported due to fear of loss of customer confidence or regulatory penalties: it costs companies on average per compromised record Large amounts of digital data can be copied at almost no cost and can be spread through the internet in very short time.

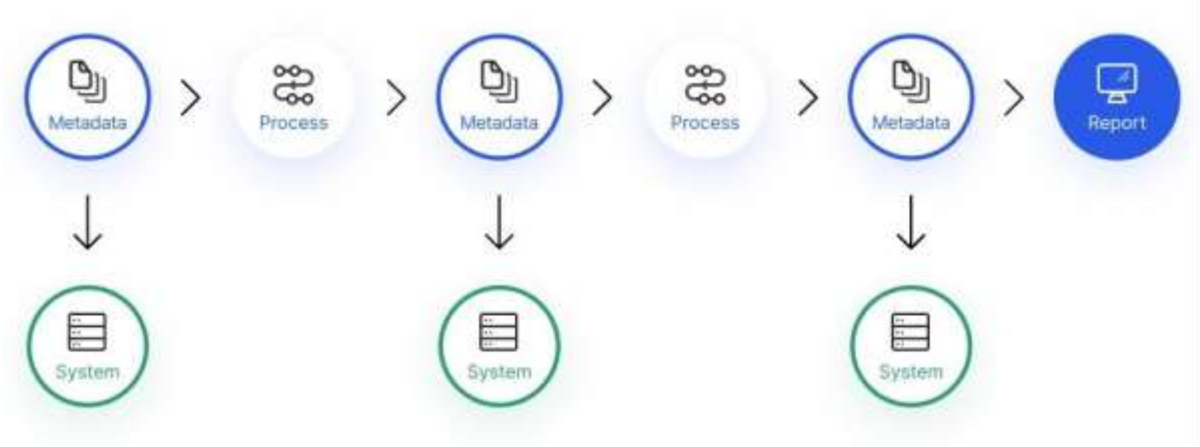
Not only companies are affected by data leakage, it is also a concern to individuals. The rise of social networks and smart phones has made the situation worse.

Data lineage uncovers the life cycle of data it aims to show the complete data flow, from start to finish.[1] Data lineage is the process of understanding, recording, and visualizing data as it flows from data sources to consumption. This includes all transformations the data underwent along the way how the data was transformed, what changed, and why.[1]

Data Lineage:

As our second contribution, we present an accountable data transfer protocol to verifiably transfer data between two entities. To deal with an untreated sender and an untreated receiver scenario associated with data transfer between two consumers; our protocols employ an interesting combination of the robust watermarking, oblivious transfer, and signature primitives.

The sending owner trusts the receiving owner to take responsibility if he should leak the document. As we consider consumers as untrusted participants in our model, a transfer involving a consumer cannot be based on a non-repudiation assumption. Therefore, whenever a document is transferred to a consumer, the sender embeds information that uniquely identifies the recipient. We call this fingerprinting. If the consumer leaks this document, it is possible to identify him with the help of the embedded information.



Data lineage allows companies to:

Track errors in data processes

Implement process changes with lower risk Perform system migrations with confidence

Combine data discovery with a comprehensive view of metadata, to create a data mapping framework

Data lineage helps users make sure their data is coming from a trusted source, has been transformed correctly, and loaded to the specified location. Data lineage plays an important role when strategic decisions rely on accurate information. If data processes aren’t tracked correctly, data becomes almost impossible, or at least very costly and time-consuming, to verify.

Data lineage focuses on validating data accuracy and consistency, by allowing users to search upstream and downstream, from source to destination, to discover anomalies and correct them.

Why is Data Lineage Important?

Just knowing the source of a particular data set is not always enough to understand its importance, perform error resolution, understand process changes, and perform system migrations and updates.[1]

Knowing who made the change, how it was updated, and the process used, improves data quality.[1] It allows data custodians to ensure the integrity and confidentiality of data is protected throughout its lifecycle.[1]

Data lineage is especially valuable in these areas:

Changing Data: Data changes over time. New ways to acquire data and accumulate data must be combined and analysed to be used by management to generate revenue. Data lineage provides tracking that makes this difficult task possible.[2]

IT Requirements: When your IT team creates a new software development process, they will need access to all data sources.[2] The comprehensive list provided by a data lineage tool saves time and money by quickly locating data sources.

Data Leakage: Data leakage is the unauthorized transmission of data from within an organization to an external destination or recipient.[3] The term can be used to describe data that is transferred electronically or physically.[3] Data leakage threats usually occur via the web and email, but can also occur via mobile data storage devices such as optical media, USB keys, and laptops.[3]

There are two types of leakage: target leakage and train-test contamination.

Target leakage occurs when your predictors include data that will not be available at the time you make predictions. It is important to think about target leakage in terms of the *timing or chronological order* that data becomes available, not merely whether a feature helps make good predictions.[4]

Train-Test Contamination A different type of leak occurs when you aren't careful to distinguish training data from validation data.[4]

Recall that validation is meant to be a measure of how the model does on data that it hasn't considered before.[4] You can corrupt this process in subtle ways if the validation data affects the pre-processing behaviour. This is sometimes called train-test contamination**.**[4]

Data Flow across Multiple:

An attacker can obtain differently watermarked versions of a document, he should not be able to create a version of the document were none of these watermarks is detectable. Further, for some watermarking schemes the input of the original document is not required for detection of watermarks. We call those watermarking schemes blind. As already stated in this definition of watermarking is very strong and its properties are not yet provided by available schemes. Although we chose this strong definition to prove the correctness of our scheme, there are existing schemes whose properties are arguably sufficient in practice such as the Cox watermarking scheme

Malicious Environments.

To take additional actions to prevent the sender from cheating, i.e. we have to fulfil the no framing property. To achieve this property, the sender divides the original document into n parts and for each part he creates two differently watermarked versions. He then transfers one of each of these two versions to the recipient. The recipient is held accountable only for the document with the parts that he received, but the sender does not know which versions that is. The probability for the sender to cheat is therefore.

DOS Attack:

A Denial-of-Service (DoS) attack is an attack meant to shut down a machine or network, making it inaccessible to its intended users. DoS attacks accomplish this by flooding the target with traffic, or sending it information that triggers a crash. In both instances, the DoS attack deprives legitimate users (i.e. employees, members, or account holders) of the service or resource they expected.

There are two general methods of DoS attacks: flooding services or crashing services. Flood attacks occur when the system receives too much traffic for the server to buffer, causing them to slow down and eventually stop.

How can you tell if a computer is experiencing a DoS attack?

While it can be difficult to separate an attack from other network connectivity errors or heavy bandwidth consumption, some characteristics may indicate an attack is underway.

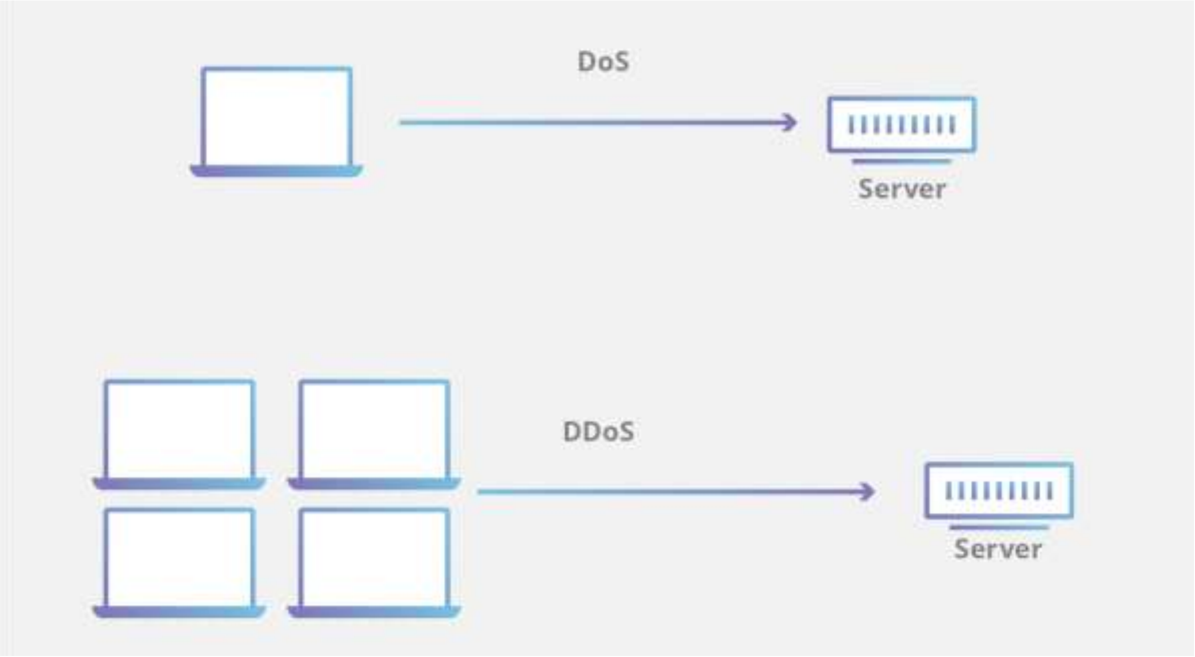
Indicators of a DoS attack include:

Atypically slow network performance such as long load times for files or websites

The inability to load a particular website such as your web property A sudden loss of connectivity across devices on the same network

Difference between a DDoS attack and a DOS attack

The distinguishing difference between DDoS and DoS is the number of connections utilized in the attack. Some DoS attacks, such as “low and slow” attacks like Slowloris, derive their power in the simplicity and minimal requirements needed to them be effective.



DoS utilizes a single connection, while a DDoS attack utilizes many sources of attack traffic, often in the form of a botnet. Generally speaking, many of the attacks are fundamentally similar and can be attempted using one more many sources of malicious traffic. Learn how Cloudflare's DDoS protection stops denial-of-service attacks.

Implementation:

Installed software: Eclipse, SQL Server Management Studio, Tomcat server

Steps required to build the project:

Step 1: Created Registration & Login Form with required validations

Step 2: User can access the system using login credentials used during registration.

Step 3: A form is created where in user can add different URLs and this data is stored in the database table named “addwebsite”

Step 4: Database connection is done using jdbc odbc driver

Step 5: After Login, User can add different URLs Under add website tab.

Step 6: Admin user can view all the details entered by user and he can view which URLs are non-secure

Step 7: can view the affected files in the system and those files can be recovered.

Conclusion:

A model for accountable data transfer across multiple entities. We define participating parties, their interrelationships and give a concrete instantiation for a data transfer protocol using a novel combination of oblivious transfer, robust watermarking and digital signatures. We prove its correctness and show that it is realizable by giving micro benchmarking results.

By presenting a general applicable framework, we introduce accountability as early as in the design phase of a data transfer infrastructure. Although system does not actively prevent data leakage, it introduces reactive accountability. Thus, it will deter malicious parties from leaking private documents and will encourage honest (but careless) parties to provide the required protection for sensitive data. system is flexible as we differentiate between trusted senders (usually owners) and untrusted senders (usually consumers).

In the case of the trusted sender, a very simple protocol with little overhead is possible. The untrusted sender requires a more complicated protocol, but the results are not based on trust assumptions and therefore they should be able to convince a neutral entity.

References

[1]. https://www.imperva.com/learn/data-security/data-lineage/

[2]. https://www.talend.com/resources/what-is-data-lineage-and-how-to-get- started/

[3]. https://www.forcepoint.com/cyber-edu/data-leakage

[4]. https://www.kaggle.com/alexisbcook/data-leakage

[5]. https://repo.ijiert.org/index.php/ijiert/article/view/994/944

[6]. https://www.paloaltonetworks.com/cyberpedia/what-is-a-denial-of-service- attack-dos

[7]. https://www.cloudflare.com/learning/ddos/glossary/denial-of-service/

[8]. Neha Belekar, R. P. Dahake 1,2Dept. of Computer Engineering, MET's Institute Of Engineering Nashik, Maharashtra, India e-ISSN: 2395 -0056 Volume: 04 Issue: 05 | May - 2017 www.irjet.net p-ISSN: 2395-0072