# #224 - The Evolution of Data Loss Prevention (DLP)

[00:00:00] **G Mark Hardy:** Hey, does data loss prevention seem like it's too big of a task to handle or the idea of classifying everything in your enterprise seem beyond your scope? I'm going to tell you a little bit about the history of DLP and what you can do today and how you can tackle that big monster. Stay tuned.

You're going to hear it right after this message.

Intro Music

[00:00:29] **G Mark Hardy:** Hello, and welcome to another episode of CISO Tradecraft, the podcast that provides you with the information, knowledge, and wisdom to be a more effective cybersecurity leader. My name is G. Mark Hardy, and today I'm going to be talking to you about data loss prevention, or DLP for short. Now, let's take a little trip back in time about how did we get to where we are now with respect to DLP?

And in fact, what are we talking about when we talk about data loss prevention? Essentially, we're looking at technologies that [00:01:00] allow us to tag or flag certain information that we want to keep on one side of a barrier from crossing over to another. In the early days that, I remember back in the 1990s, we had something that we nicknamed a dirty word list.

And this is something we had in the military when we had classified networks and unclassified networks. Now, there were opportunities to move information back and forth, but of course, you wanted to make sure. You, reflected, it was a star property from the Bell Lapadula model for those of us who took a CISSP many years ago, where it's okay to, read information up, but you can't move classified information down.

And so the idea was, is that if you knew of certain terms or project names or anything else that would, Indicate that something were sensitive or inappropriate to be put in the unclassified world that your dirty word list would then pick up on a project name or word secret or confidential or something like that and flag it so you could go ahead and say, Hey, maybe this really shouldn't be doing it.

So that was probably the [00:02:00] earliest instantiation that I saw. And that was, it was almost 30 years ago. And what that happened is, you had one network here that was operating at one level system high. Another network operating at a lower level and this sat in the middle because there was only one way to get from here to here.

Pretty simplistic model under the circumstances. That's all we had were simplistic networks and life worked pretty well. Now, a little bit later as we go ahead, we find out that as risks increase, we have not only accidental exposure or insider threats, but even external attacks. And we have a lot more external attacks today than we ever did before.

but We moved on, let's say, to the early DLP, what we might call simple content filtering, and this is probably from around early 2000s and included endpoint protection. So then what was happening is we're getting regulatory requirements, requirements like HIPAA, Gramm Leach Bliley Act, and Sarbanes Oxley.

And this is pushing companies to go ahead and protect and restrict [00:03:00] their sensitive data from moving around. Now, these initial DLP solutions were focused pretty much on content filtering and endpoint based controls, similar to what I had explained before, but now with a little bit more sophistication.

Typically, what you'd find then is being able, instead of just looking for a word, you could write a regex, a regular expression. And regex matching allows one to go ahead and look for things such as credit card numbers, three digits, a hyphen. Or I'm sorry, social security numbers, three digits, a hyphen, two digits, a hyphen, four more digits, or credit card numbers, which would be four groups of four, 16 numbers, maybe email addresses, anything that might be considered to be sensitive, PII, PCI, if we're looking at credit card data and things such as that.

And so those are the early usage where you'd look to match something in an email or transferred file because you're looking for words or phrases or something that would suggest that it's confidential or proprietary. The difficulty is though, is you're going to end up with a high false [00:04:00] positive rate.

Why? Because it has no comprehension whatsoever as to the context. So if I have something where I was saying things such as, instead of classifying a document as confidential, I would say that, last week we held a confidential meeting with our client, and it went really well. that would get flagged as a

potential false positive, even though you didn't disclose anything that you needed to protect.

And so on the keyword matching, it came up a little bit short, but it was better than nothing. And on the endpoints, what would happen is you're looking at data movement. So what happens when I stick a USB device in there? Someone's going to exfil data. I can identify that and block USBs from working. Of course, back then we had the old PS two mice and keyboards, which is a different port.

And today, where if you have an external mouse or keyboard, it is pretty much guaranteed to be a USB connection. you block USB, then you can't operate the keyboard, you can't operate the mouse. That becomes a little bit of a problem. And you had tools like device lock, you could disable that. But [00:05:00] the problem with that might be that, you disable functionality.

Because if you don't discriminate against a USB data device versus a keyboard or a mouse and then a little bit later, of course, you have, for those of us who are familiar with it, the concept of a rubber ducky, which was a USB that declared itself to be a keyboard and then it would allow you to queue a keystroke injection.

that's a little bit later, but let's not even worry about that. But again, high false positives is a problem due to lack of context. Didn't scale very well when you increase the data volume because now if you're trying to move things at a much faster speed, your processor might not be able to keep up.

And a lot of times you couldn't support complex file formats, or even things like structured data. Or what we used to do is it was looking for your .Doc and would review that, or you say you can't send a .Zip. So what do you do? You rename the file extension. Instead of zip, it's zap. And then it goes, oh, that's not interesting.

Off you go. And these are almost trivial. To go ahead [00:06:00] and get around now, the next phase, probably let's do it a little five year increments. Somewhere around 2005 to 2010. We moved on to advanced content inspection and policy enforcement. So now what's happening is we in the context of data breaches.

Remember the T. J. X. Breach that took place back in 2007 there's more regulations coming and so as a result we have to up our game to be able to defend against the loss of sensitive information. So some of those technical

approaches would include things such as fingerprinting where you could create a unique hash of a sensitive document.

Some structured data set, et cetera, so that any attempt to move that block of data, you get a hash match. You have a signature match and you say, Nope, can't do it. Not allowing it. And so now I could use and say, I have this database. It has all my client data. Here are all the files that pertain to customer X.

I record those fingerprints at any time and attachment. It's pretty trivial and pretty quickly to produce a hash [00:07:00] of a document, no matter how big it is. It goes quickly. And say, yep, that's a direct match. Now, of course, you're already thinking what's going on here. We have to say, how do I know that this is sensitive?

How do I tell somebody that you need to add this file to the database? I have to have the concept of the classification engine. I want to be able to categorize data, and it could be public data, which is the lowest level. It could be internal, it could be confidential, or it could be highly confidential.

That sounds a little bit like purviews defaults. we're going to get to that in a little bit. And so now what happens is you can add tools like network based data loss prevention. It's going to monitor the network traffic. Now, this is pre Snowden, so an awful lot of stuff was going HTTP. Or even file transfer protocol, FTP, SMTP, mail going back.

And this stuff was all in the clear. And so as a result, it was pretty trivial to go ahead and grab these things, inspect them, and sit in the middle, and tools like a WebSense would go ahead and pioneer this network based filtering for [00:08:00] sensitive content. Now, what about a scanned document? there's a solution for that and that was being able to look at images and say, we could build an OCR engine and that OCR optical character recognition engine could then go ahead and translate that into ASCII text, obviously a lot less space than it would require to score that image.

But now I can run operations on it and I can go back and use my matching algorithm before to say three digits, a hyphen, two digits, a hyphen, four digits. Looks like there might be a social security number in that scan document. We probably want to do something about that. And now what we can do is we can start to centralize our policy management.

We can define rules for different types of data. And with a tool that works enterprise wide, we can enforce things like block all PII from being sent over

email. Alrighty, again, that works great when your information and your data is structured. But in unstructured data, it's a little bit harder.

Again, if you have to update your classification [00:09:00] rules, there's a lot of maintenance involved. If you miss something, then all that stuff keeps going by. And of course, once it's passed your filters, it's lossed. And with a lot of frequent blocking false positives, the users weren't all that thrilled about that.

Now, let's go to another phase where we would have a unified DLP, let's say the first half of the. The teens, the 2010s or so. Now we're going to integrate with the same endpoint agents. And what we have is a cloud adoption. In mobile workforces, this change in our behavior, instead of just client server, here's my endpoint, here's my server, and then maybe server to server.

Now, what happens? I've got mobile workforces. They're all over the place. They're on laptops. They're coming in from a lot of places. I have cloud adoption where I'm moving things up into Amazon and Microsoft and Google and all the other cloud services that are out there. And that's a problem because now I'm not running like a perimeter type of a defense.

I don't have a perimeter anymore. So where do I go ahead and put these DLP things? [00:10:00] it turns out that Some vendors started integrating endpoint and network and cloud DLP solutions. So you had a single pane of glass. You could monitor endpoints, monitor your networks, monitor cloud environments. And then you could take a look at these unified DLP platforms and determine to say, it looks like we've got some information that might be moving from the cloud to an endpoint, or maybe going ahead and moving across networks.

And we could go ahead and identify it. What we came up with is a concept for behavioral analytics. And U E B A, if you remember that alpha Algorithm, acronym, sorry about that. User and Entity Behavior Analytics, UEBA. You could allow DLP to detect anomalies based on user activity. For example, if a person is normally doing a certain amount of transaction volume and all of a sudden there's a huge peak, like they're exporting a ton of files or diagrams or internal things, you could flag that as a potential insider threat.

Also, you could start to look at data sensitivity, not based on a pure list of [00:11:00] classifications and dirty words, if you will remember the back and the old days. But now look at the who, the what, the when, the where, the how the data were accessed. Provide some context so that if you said, I've got sensitive data transfer with internal systems, that should be okay because it's not going out.

Even though it's going across the network, it's going from one end point to another. The destination is still internal. And the route is still internal is compared to something that's going external, which we say, that violates this rule. We ought to block it. Now we could go ahead with the seams and integrate this DLP data so that we could correlate it and do some threat analysis.

So things like Splunk and ArcSight, they would ingest the DLP alerts. And now you'd have deeper insights as to where potential problems might be. Also, another innovation that really started to come of age then is tokenizing. Data upon detection of sensitive information. So you might say, I do need to move it from A to B.

And A is my Los Angeles office, B is my New York office. They're both [00:12:00] internal, but I've got to go across a public internet. So how about we go ahead and tokenize it so we agree on something that the client name, if that's sensitive or some project is sensitive, we just substitute something for it. And then we move it across and then we re substitute for the second user what was really there.

Now you could go full bore and encrypt everything, which would be better in my opinion, but then you have to manage keys and things such as that. If you do it right, it works pretty well. It's now getting more complex, and with this breadth of coverage, these multiple different platforms, not only are you going to have to worry about performance issues, how am I going to keep up with this volume, but you might have a whole lot of alerts.

And as a result, you're going to create alert fatigue. It eventually becomes like the boy who cried wolf. You stop paying attention to it. The next phase, if you will, of DLP got into cloud and software as a service using tools like a CASB, a cloud access security broker. And now this is a companion.

To traditional DLP and software as a service and the [00:13:00] increased use of that back then says, Hey, we're going to go ahead and push everything up into the cloud tools like what's now called Microsoft 365, but Oh, 365 coming out of age back then. And Hey, I'm moving stuff up in the cloud. It's moving out there.

It's going to be in my one drive, et cetera. You have to go ahead and account for that. Now with CASB, what I could do is take this cloud access security broker tool, inspect data in my software as a service platforms. Now, if I'm using something like O365, now Microsoft 365 or Salesforce, I go, all right, that should be there, that should not be there.

And Microsoft Defender started coming in line to integrate DLP with your cloud app security, so that we could go ahead and potentially flag where problems might be. In addition, vendors came up with plugins, APIs, so you could have a direct API connection to a cloud platform, and you could do real time data inspection and blocking.

Those like Netscope, for example, that would work with Dropbox and G Suite, and even inline proxy DLP where I could [00:14:00] leverage. proxy based inspection to monitor and block data flows to and from cloud services. So basically sit there in the middle. Zscaler at proxy based DLP for secure web gateways.

You could have identity aware DLPs. You could integrate with identity platforms that allow for kind of dynamic policy adjustments based on the user roles and permissions. Forcepoint. Had a tool that would allow you to adjust your policies based on active directory groups. About that time, I was working on a startup and I remember that they had a blue coat and that firewall was set up to go ahead and essentially provide a man in the middle where it would serve up security certificates that were rolled there.

So what happened is you said, Hey, what's my challenge? Everybody's encrypting everything. How do I view the encrypted traffic? If I control the certificate that your machine trusts, then what I can happen is when you ask for HTTPS and you do the TLS protocol, that protocol exchange is not with the ultimate server.

What happens is [00:15:00] your device, your man in the middle, so to speak, that the blue coat was doing, As it would say, Hey, I'll accept that call and I will go ahead and do the four way handshake with your TLS. Now, at that point, I have a secure connection, but that secure connection is to my trusted man in the middle, which would then decrypt it, allow me for inspection, check for DLP, then re encrypt it, because it's going to go talk to the ultimate destination.

So what you have in the middle is an instantiation of unencrypted data, which allows you to then go ahead and inspect everything, even though it's HTTPS. And, with Ed Snowden, everybody else say, we ought to encrypt everything. The amount of HTTP traffic and all the stuff in the clear really went down as everybody got religion on encryption.

That could create some performance bottlenecks because you've got to do this real time cloud inspection. People don't want to go ahead and wait for things. You've got this on prem and cloud. How do I coordinate these things and know

that, this cloud is mine, but that cloud's not mine. And this is okay, but that's not okay.[00:16:00]

And then the encryption, tokenization, things such that gets rather difficult. Finally, if you will, our current phase, we're now getting into AI and machine learning DLP. Now we can adapt and automate our protection because with this massive amount of unstructured data out there and potentially increased insider threat, we've got to have a more intelligent.

Automated protection. What can we do about that? One is, use machine learning. And machine learning could actually be used for data classification. So my AI driven classifiers can improve my accuracy, reduce my false positives, and allow me to go ahead and have a higher fidelity for my process.

Microsoft Purview DLP would use machine language to distinguish between sensitive and non sensitive data because it can ingest a lot of the information and infer a context. Natural Language Processing, or NLP, allows context based understanding of human language and documents, communications, so it could figure it out.

So tools like Forcepoint would look at that and say, this looks like confidential contract language. [00:17:00] It should probably be protected even though it wasn't correctly marked. And then have that policy tune itself, allowing the machine learning to adjust automatically based on user behavior and content.

And therefore, tools like Proofpoint, Forcepoint, you can allow that real time so you adapt to your user's behavior and you don't complete the continuous cycle of false positive, Once it's been told or trained that, hey, this part's okay, off it goes. And then finally, zero trust integration.

As we make DLP a key component of zero trust strategies, we can tie data protection to identity, location, device health, all the things to ensure that if we're going to encrypt and authenticate everything, we should be okay. Zscaler, NetScope will work pretty well on that. But that's a lot of complexity trying to tune these machine language based models and a lot of high compute requirements.

If you're going to do real time data classification. And of course, think about it, there's a privacy concern. If you're going to be doing deep inspection of communications, something that dives [00:18:00] deep into all of your stored documents, how do you feel about that? Is that a privacy issue? maybe. So

today we're seeing a shift from standalone DLP to essentially integrated data protection.

We use Secure Access Service Edge or SASE platforms. Potentially receive more homomorphic encryption, meaning I can work on the encrypted data without decrypting it. Now, it's interesting idea and it's been proposed for a long time, but there's a couple of functional instances out there. Being a crypto guy, I might learn a little more about that, because I haven't studied all that much, be able to say, hey, I can produce a result, operating on two encrypted inputs, create an encrypted output.

It's as if I were working on. The clear text, but I still get the right answer. Interesting concept, but I can see how mathematically you can do it. Maybe we'll save that for another episode. Agentless DLP models. That way you don't have any footprint on your endpoint. And you can go ahead and improve your performance because it's going to go ahead and just look [00:19:00] at stuff as it goes by.

And focus on things like data sovereignty, cross border data controls. If we have GDPR and other regulatory requirements that put some pressure on us, we're going to find out that's a bit of an issue. I'm working with a client this week on how do we go ahead and do data classification and use Microsoft Purview.

Now I'm just going to do a quick overview because I'm not going to dive deep into it, but I just want to use this as an example of things you can do currently. And Microsoft Purview, which is plugged into the 365 environment will give you visibility to all of your data sets, let you manage it across your environment, help you secure your data throughout its lifecycle between when it's created, transmitted, modified, archived, and eventually destroyed, and help with risk and regulatory requirements.

So one of the course, the first things is if you're going to look at doing DLP, where is my data and what is it? And how do I access it? And so if you have everything. In a known location, [00:20:00] let's say OneDrive and or SharePoint. But you could also have Box and Dropbox and other services and things like that.

You really need to build a data map. And using a tool like Purview, they say we can help you do that. Take this metadata across all these other environments, whether you're hybrid, on prem, multi cloud, software as a service, and be able to say, hey, we can see where all this is, and we can ingest all that and create a data map.

Okay, then what you want to be able to do is make your data discoverable so you can maximize its value. To do you create a data catalog. And now the data owners or stewards of data can curate their data assets and say, Hey, here's what it is. So if you're a consumer, I can search for things and say, that's where I need to find it.

Again, not a big deal in a small, simple organization. When you get large and complex, it could be difficult. Then we want to have policies for accessing the data, moving it, perhaps even sharing it. And again, data policies that you specify these. So you could go ahead and say, I want to create and [00:21:00] apply different policies based on who's accessing it.

If it's DevOps and you're looking at code, that's probably okay. The data owner, that's probably okay. somebody over in accounting or somebody that's all in the assembly floor probably should not be looking at source code. And so be able to go ahead and limit that. And so to do that, we're going to want to secure our data and we're going to secure it throughout the life cycle, whether it's on an app, in a cloud, in a device, wherever it happens to be.

Key to this, and this is the hard part that I'm facing right now, is discovering and labeling everything. Because I could start labeling all new information tomorrow, but what about terabytes of data? That have already there. Do I have to go back there and pick through all these files that I may never ever care to look at again.

And so it seemed like a tremendous waste of time. So the idea is, having tools that could go ahead and software development kit where you can place things on there to say, Hey, look at all my old client files [00:22:00] and here are the types of things that would be sensitive. pricing. How do we go ahead and have a proprietary method for doing the engagement?

But standard boilerplate stuff that would be in our legal terms and conditions may not be so insensitive. So then go ahead and say, let AI, let your automation, let it go scrounge all through your data. So you don't have the tedium of doing that. Then we want to deploy our DLP policies to restrict our data leakage.

And now data loss prevention is the capability of purview that lets you know when things are going out. Now I'm using that right now and I've got it in a notification mode rather than an actual interdict mode. And I get a lot of false positives. I gotta tell you that most of them are with our chief financial officer, warning credit card number going across, warning bank account information going across.

And I look at the source item and I'm like, Yep, that's a CFO. Now, yes, I'm going to filter that out at some point, but I want to see everything that it does. So I get a little bit of noise. But as a CISO or the person you designate, you want to understand what this thing is [00:23:00] capable and not capable of.

It's a little bit easier to take too much information and trim it down with filters than it is to start with not enough and say, how do I get that stuff that I'm missing? And so now what they're also as tools available for insider risk management, being able to use machine learning can look for potential insider risks.

So you could say, Hey, this just doesn't look like what a normal person ought to be doing. A big challenge that most of us face or compliance and compliance represents a tax, if you will, a restriction, a constraint on your operations, that says you have to manage things in a certain way. Why? because The regulation says we don't want to disclose personal identifiable information, protected health information, payment card information, all these other types of three letter acronyms beginning with P and ending with I, that say that, yeah, this has to be protected because not only do you run the risk of the embarrassment of disclosing something, But you may have financial consequences for violating [00:24:00] that as well, as the reputational damage that could occur when somebody says that you just had a major data breach and things such as that.

So by being able to build in classification and governance at scale and recognize patterns, these are the types of machine learning types of capabilities that you're looking for with your tool. And then ultimately, as I said, you have to keep core business records. When do I dispose of data? Information is not always an asset.

It can become a liability. And in a situation where you've kept data far too long, and then something comes along where somebody's doing discovery in a court case, and they ask for everything. not only do you have to produce that, because it's a court order. But they're going to charge you, probably, to say, here, we spent hundreds of hours reviewing all the data you sent us, and we charged by the hour.

Thank you very much. Also, there might be things that took place in the past. It could create a potential liability today, because let's face it, when somebody goes ahead and goes to court, they don't try to put your best foot forward. They're going to take everything, as much [00:25:00] as they can, I would argue, potentially out of context, to try to use it against you.

But if you had an absolutely strict Retention policy that said interns, when they leave after six months, all the data is gone. Employees, after two years, all the data is gone. Executives, three years, all the data is gone. Except if you have things like a FINRA requirement or some other regulatory requirement for it.

Now you enforce that and you do it over and over again. Now you get served with a subpoena. We said, hey, you had an intern that worked with you four years ago and this person is now facing some harassment suit and we want everything that person ever said four years ago to try to either build a case or defend a case.

At which point, you just hold up a copy of your data retention policy. You say, I got logs, but since such and such a date, when you prove this policy, we've been deleting things on time on schedule. There is no business purpose to keep this person's stuff after six months. It's now been three or four years.

I cannot give you anything. And that'll hold up in court. they say, Judge, they're not giving us anything. Why aren't you giving them anything? Our policy, which [00:26:00] we've enforced for years, says we delete everything, and we've been enforcing it. And if you have those logs, they're like, yeah, we got. Also, make sure that if you do have logs, that they're unalterable, that you can go ahead and prove that.

Who knows, maybe stick them on a blockchain again, episode for another day. And so now if you have e-discovery, that's a tool that allows you to go ahead and preserve and collect and analyze and review and export content that are based on a particular investigation. So now, instead of having to go through and manually pick through every single email, every single file, every correspondence, let the AI tools do it.

Let some intelligent machine learning do some deep indexing, email threads. And the near duplication detection. Because one of the things I found is that when you do discovery, if I send an email and then you send a reply, and then I send an email and a reply and send my new end code back 10 times, that's 10 different emails I want.

Now you're supposed to print them all out and even have them by 11 depending on, certain size font. They might take them electronically, but the whole idea is, that you want to go ahead and say, look, this is all one thread. The 10th message contains [00:27:00] 987654321. It is unaltered. I can prove it was unaltered.

I'm just going to give you the one. If you want to start pulling strings after a while, they realize, yep, they got their act together and the court's going to leave you alone. And then forensic investigations, being able to go ahead and have audit ready reporting, insights and things such as that.

This is going to require some specific information, but now I can look at user activity. When all of a sudden I have a peak in data access that's high bandwidth, what's going on? what's happening and things that might cause an audit to occur. And now I can preserve audit logs. And, for example, Microsoft Purview now lets you keep them for up to 10 years.

It's compared to your normal data, which is going to be about 30 days, unless you're on academic license, when I think it's about seven. And then let's look for violations, potentially. If somebody is violating regulatory compliance or business contact or sending inappropriate communications, don't wait for the lawsuit.

Stop it and quash it immediately. And then go ahead and ensure that you're remaining [00:28:00] compliant so that if you're dealing with things like SEC or FINRA and you have somebody who's making promises that they can't deliver in terms of investment results, you can spot that very quickly and shut it down and not just hide the evidence, but maybe have a follow up to say, hey, this is compliance.

Yesterday, we received an email from one of our employees who said, I can get you 10 percent on a risk free investment, A, that is not authorized. B, we do not authorize that communication. please disregard it and see that person's no longer working here. But again, build yourself an audit record so you can go ahead and protect yourself.

And then we want to make sure, ultimately, that you're tracking your compliance effectiveness. How well am I doing? Can I continuously look at my compliance efforts? Have I reduced risk? Because ultimately that's what I'm trying to do for our organization, our executives, is reduce the uncertainty of bad things happening, so that I can go ahead and meet the requirements that are out there.

So these are the types of tools that I can do with DLP, and I gain a much better capability for my enterprise. So [00:29:00] in summary, if we think about data loss prevention, it's a way of ensuring that what goes out, around, through, or past some barrier, some perimeter, some defined point, meets a predefined set of requirements.

Of course, you have to predefine those sets of requirements. And some tools like Purview give you a head start. They will give you four different classifications. You can start putting different tools in those buckets. information in there, let it automatically scan, let it build up that case for you and then proceed going forward.

It looks daunting. It really is. And so for a long time, I was like, do I really want to open that can of worms? But as you move forward, you realize that as a CISO, as a security leader, you're going to have to open some cans of worms. Why? Because they're going to start to rot and smell and cause danger for your organization if you don't tackle the project.

So if you've been holding off on DLP, if you've been holding off on data classification, if somehow you felt this isn't for me, or this is too big, or I don't have the staff, or I don't have the bandwidth, today's your day because we've got AI, we've got all these capabilities [00:30:00] with machine learning. And all those tasks that looked nearly impossible to accomplish as human beings can be done pretty simply with these tools.

So hopefully you found this episode useful. If so, share your insights with others. Let them know that you're getting them here at CISO Tradecraft. Follow us on LinkedIn if you're not already following us. We have more than podcasts. Make sure you watch us on YouTube if you're not subscribed there. And Let other people know the source of your CISO tradecraft so you can help them in their career paths as well.

Thank you very much for listening or watching. This is your host, G. Mark Hardy. Until next time, stay safe out there.