EPISODE 1459

[INTRODUCTION]

[00:00:00] JM: Data loss can cause sensitive information to escape from platforms such as Slack, Google Drive and a multitude of other systems. Data loss prevention allows for monitoring and alerting around potential data losses that can cause issues with HIPAA, PCI and PHI compliance. Yasir Ali is the CEO of Polymer, a data loss prevention platform that detects potential data losses and allows for companies to respond. Yasir joins the show to talk about the engineering of Polymer and the market for data loss prevention.

If you're interested in sponsoring Software Engineering Daily, we reach over 250,000 engineers monthly. Send us an email, jeff@softwareengineering daily.com or sponsor@softwareengineeringdaily.com.

[INTERVIEW]

[00:00:41] JM: Yasir, welcome to the show.

[00:00:42] YA: Thank you, Jeff.

[00:00:42] JM: You had significant experience in consulting and other areas before you started Polymer. Can you describe how your work informed what you've built with Polymer?

[00:00:57] YA: Yeah. I have a computer science background, and I was in enterprise tech and architecture, working for financial services, and solving some cloud migration, and kind of MDM data warehouse type problems. One of the issues which I saw, which is not just endemic to technology side of the business, but also in general in large organizations is a silo effect. Where security teams necessarily do not talk to the data governance team or the data program team would not be talking to the business teams necessarily. Where different teams are trying to solve the similar problem and where essentially talking across each other or not solving the same problem.

The problem which came about in my experience was data management and data governance, especially around data privacy, which became a big thing, post CCAR, which the Fed implemented in 2010 for large financial organizations. Then GDPR, implemented by the EU about like six years ago.

Saw the issues around data governance, which were being solved in completely different ways. One side by security teams and one side by data governance team, but not being effective. Polymer essentially tries to combine those common business outcomes of data security, data governance, which we look at it through the same lens, rather than two different sides of the equation, which a lot of businesses are still trapped under.

[00:02:33] JM: Give an overview of what you're building with Polymer, and then we'll get into some of the engineering challenges.

[00:02:40] YA: Sure. We are essentially in a nutshell, it's a data loss prevention software. The old construct of data loss prevention used to be for on-prem solutions, where it was on the network. It was on your email servers. Any data traffic would get a cursory inspection, whether it matches a credit card pattern using regular expressions. It will contain anything matching a security pattern. The firewall or your data loss prevention adopter will block it from leaving the premises, your on-prem premises.

With the cloud and the cloud workflows, this problem has become quite unmanageable. Whereas, especially if you are using third-party SaaS applications, you can think about Google Drive, Slack, Zoom, so and so forth. The issues of data loss prevention become even more magnified. Because these products, these platforms are meant, designed to have a network effect where you want the maximum number of people using them and transacting and sharing information. The limitation in these platforms is the IM policies in general at an organization level, the way they've constructed allows you access to an application. But once you're inside an application, you can pretty much share nearly really pretty much anything. That's kind of where we come in as a data loss prevention product for third party SaaS apps, which can monitor, and remediate and stop some of this proliferation of sensitive data sprawl, both internally and externally for organizations.

[00:04:23] JM: There's a wide variety of places where companies are storing data. A lot of them have a gradient of permissions and a lot of that data is going to be available to maybe in some cases, 1000s of employees within the company. These might be documents in Google Drive or tickets in Zendesk. A lot of this can be sensitive data. How can you differentiate between uses of that data that are valid and friendly to the company versus uses of the data that might be categorized as data loss?

[00:05:09] YA: Yeah. I all starts with the constructing a framework, a new way to look at how I am is implemented. IM Construct has been around for about 40 years, where it's based on access to a folder, access to an environment and access to an application. When we think about unstructured data, that is exactly as you put it. It becomes almost impossible to differentiate between one document and another at a document level, what is considered sensitive or not. Our approach is to essentially do a deep inspection of the data contents itself using NLP and other methods of data inspection to be able to extract out entities within these unstructured data sets.

We built a framework to allow mapping of access levels to entities for users or groups of users within an organization. Example, HR team should be able to see employee data. Finance teams should be able to see financial information sets, but it cannot happen vice versa, for example. That's kind of where we think of an IM policy, which is a little bit lower level than where current IM policies reside is the starting point around being able to very granularly define access to datasets even within a document itself.

[00:06:37] JM: Got you. Can you give me a little bit more context on how you're analyzing these documents?

[00:06:43] YA: There's multiple ways. In essence, its natural language processing, which allows for an easy feature engineering by the organization to purpose fit for their uses. Every organization has different ways of looking at similar kinds of data. For example, in HIPAA, one covered entity from another would be looking at HIPAA data much differently. We need a way to be able to allow organizations to feature engineer, feature set in terms of tweak the base NLP model in designing or differentiating between what's sensitive for them. That's number one. That allows us, frankly, high level of accuracy. Then obviously, if there are images involved, then there needs to be dynamic OCR that needs to be done. If there are other types of documents, obviously, we use regular expressions. We can even go as far as doing lookup values against a known data set of sensitive items if needed. There are multiple tricks in the back to be able to analyze and differentiate what is sensitive within these datasets.

[00:07:48] JM: How do you integrate with the wide cardinality of available data sources. You got a range of things from Slack, to GitHub to Zendesk, and you have to normalize all that data and suit it to your NLP engine. Just give me some description of what that integration and normalization looks like.

[00:08:14] YA: Yeah. It all comes down to having an abstraction layer, which does not discriminate where the data is coming from. That abstraction layer, basically, all it's looking for is a data object and data object can be an image, Word document, 2004 Microsoft Excel. It could be structured data sets, data frames for example. As long as there is an incoming data coming in into this abstraction layer, it needs to know how to handle it and push it back into the requesting API. That abstraction layer took us a while to build, but once that was done, connecting into individual different SaaS platforms, that is actually now become fairly regimented, where the beauty of the SAS world is, there is app store so there is some known APIs. There is known hooks and points to be able to connect into, or you can kind of build your own if need be. But there is a far more popular SaaS platform.

There is enough availability of booking into triggers, hooking into events, logs into these systems. Once that piece of it is done at an integration level, but usually that is separated out from this abstraction layer, which does kind of handles this data set. Down the road, this could be from non-SaaS platform also. I mean, we've done on-prem installs, where some of these events are coming in from on-prem systems also and the system doesn't care where the data is coming from as long as it knows there's a data it can handle, data type it can handle.

[00:09:47] JM: The system of analyzing a document consists of NLP and various machine learning techniques. Can you just talk about how you've trained the machine learning model to identify data that is of some level of risk?

[00:10:08] YA: This is actually for the two main industries we focus on, which is the financial services and healthcare. **[Inaudible 00:10:17]**, there is enough experience in the team to be able to understand what is patient data, what is financial information sets. We've done this in the past many times. That was not the most difficult part in terms of understanding of a name entity versus an address entity versus a credit card entity or Bank of America account number entity, for example. Those things have become more and more trivial to be able to extract out. The complexity happens around the false positive or reducing the false positives. The issue is, that in many organizations or HIPAA is a classic example, where it's only considered to be sensitive if different elements are combining together. Patient date of birth by itself is not sensitive, but if I can tie the patient date of birth with a patient name and the medical ID, then it's a linked data, which then starts creating. You can basically go back to the original patient-based on two or three different data points, find in the document, or within a paragraph or within a line item, for example.

That's going to where a lot of work was put in, when we were kind of launching this business. Think about how do we make this clustering or solving the nearness problem. That's where we have allowed our clients to be able to come in, and with a no code methodology to be able to like drag and drop different pieces of entities. If only certain aspects of entities are found, or certain number of entities have found, which ties into a user or a patient or a customer, then only does it trigger an alert. Which is kind of what the challenge is, versus a dumbed down extraction of entities, where you will be just overwhelmed by false positives. That is, frankly been a challenge with current DLP solutions so far.

JM: What's your engineering stack look like for ingesting this data and running machine learning jobs over it?

[00:12:18] YA: It's a LAMP stack. Its AWS based. We use a lot of Lambda based processing. It needs to be highly scalable, because we don't know when our event logs are going to happen. People could be going crazy on Slack in a 5000 or 10,000-person organization, a 100,000-personal organization for that matter in any given point in time. How do you kind of handle all that traffic at one time? Every single document, every single event triggers a Lambda process, which does its job, and it could it could spawn up and be able to then essentially push back the results, which are then queued and put back into the SaaS environment or the policy engine, depending on what the event is.

That Lambda-based processing has been a great savior for us. We're constantly looking to make that more efficient. We know that after 10,000, Lambda processes running concurrently, it becomes a little bit dicey, so we figured out ways of kind of like handling that, how do we look into other ways of scaling that beyond the 10,000 processes. But the good thing is, every customer we hook into have their own environment. We have 10,000 buffers at any given second per customer. It's not too bad. We won't run into that too much, but we can see that happening down the road.

[00:13:37] JM: As the Slack messages are being produced within an organization, do you have hooks into Slack that's constantly analyzing messages for their sensitivity, or is it more of like a batch offline process?

[00:13:56] YA: No, it's real time. That is a requirement by many of our customers in terms of reducing the surface area of sensitive data with these platforms. There is an inline check that happens. If it doesn't pass the muster in the inline check, it goes back to our data layer or data transaction layer after that. If there is a file, a document, then typically we will push it back anyways for deeper analysis, because the inline check is very basic. We try to optimize a little bit in terms of not pushing everything back to the Lambda process if we don't need to, but it's all pretty much near real time, as I would like to call it.

[00:14:30] JM: Gotcha. Let's just categorize a data loss event. Can you explain how you actually classify something as being a data loss event?

[00:14:42] YA: Yeah. We break down the SaaS, the world into four main areas of focus for us right now. Storage systems, like Box, Dropbox, Google Drive, their ticketing systems, like a ticket in email to me on a Zendesk, Gmail, Microsoft 365. Then there is a code basis, so GitHub, GitLab, etc. Fourthly, it's essentially chat platforms like Slack and stuff like that. All those different four categories of environments have a different form of data loss.

In a storage environment, in Google Drive, let's say, data loss could happen if someone shares a link to a file having sensitive data to someone external, who is not authorized to see X amount of entities within a patient data, for example. If that happens, in this example, the machine might not necessarily look into redacting anything, because this is a severe event. It will just go out and automatically within like a few minutes go kill those links to those external people in Google Drive, within a certain time of SLA. In Slack message. If there are sensitive files shared in a public channel, the machine will go out and redact anything, which is considered sensitive.

Within that unstructured data set, within a few minutes of that data being there, sometimes in seconds. Also, if it's a small data set.

One level, we want to minimize the amount of footprint. On the other level, we don't want to increase too much operational friction into this platform. We try to stay rather than having users not have any context, in terms of what happened, they can still see the redacted version of the file, and then be able to unredact it all by virtue by clicking a button, for example, within Slack or in Google based on the policy, they can request in admin that they want to send it to someone as a link, rather than through email or some other methods.

In GitHub, it could mean a code base having keys and there are certain amounts of sensitive data in there. Whatever is defined by the organization, that code base can be stopped from being checked in, or an alert can go out to the admin. There are multiple ways the machine can be tweaked, it could be run in warning mode, tracking mode or full remediation mode. It depends on platform-to-platform a little bit. But the net effect needs to be reducing the blast areas, reducing the amount of time the data is exposed to unauthorized people. Those are the two main inputs into our concept of a risk score, which essentially is – the goal is to minimize the risk score over time as quickly as possible.

[00:17:14] JM: Tell me more about calculating a risk score.

[00:17:17] YA: Yeah. Risk score for us is essentially, it's a multiple of amount of sensitive data, time it's left unsecured, time it will open to an authorized people, times the number of people who have access to it. Logically, that's usually the equation that it works out. In every platform, it's a little bit different. But the whole goal is that what we saw in the market, especially in security solutions, in general, especially in data governance solutions, there is this kind of – I won't say laziness, just a function of the way the market is evolving to. It's based on this human loop in the middle, which talks about like, "Okay. There are all these alerts, every alert is equal to the next alert itself." We want to be able to differentiate between a document having your employee salary information, having a higher risk than potentially having a document, which has maybe an NDA built into it, for examples.

There needs to be some way to sign differently to different sets of documents, depending on what contents they have, and who they get access to can really have a lasting impact to the organization. That is the goal here and our future kind of state, the way we want to play with the risk score is, if the risk score of the organization goes up, the remediation pieces within the platform can get tightened up automatically. We want to kind of get to a fully autonomous security landscape and solution over time. This is a kind of early goings towards that regard. Right now, a risk score is used to essentially measure performance, and then that could change some dials based on, "Okay. Rather than having a document exposed for 10 minutes, we can tighten it up or loosen it up if our risk code is kind of at a low right now.

[00:19:04] JM: Gotcha. Would different organizations want to calibrate different risk scores or different response strategies for different risk scores.

[00:19:14] YA: Of course. For some people, sending a public link based on their governance or privacy policy, we could have a risk score of 20 for a document, which is, let's say public versus a restricted link to individual emails could have a risk score of two, so a factor of 10, let's say in this example. Every organization has a different risk tolerance or different methods into calculating this thing. We do work with our clients to be able to compute that and set it up day one.

[00:19:42] JM: Speaking of different domains, can you give a description for how different verticals like health care or financial services might use Polymer in different ways?

[00:19:56] YA: Yeah. If you look into the statistics, 80% to 90% of security breaches happen because of human risk. One of our goals is this whole feedback loop we want to incorporate where employees can see the results their actions are causing. We are very big on sending end of day reports, alerts within Slack and other methods to be able to say, "This is the risk you have created, or these are the files you shared, or did something which was unsafe." We want to build a culture of trust and privacy, number one. That is the end goal, because it doesn't have to be a function of your security team, or your data governance team or compliance team to manage data privacy, and data security. It needs to be everyone's job; the whole organization needs to be part of it. That's the only way this can succeed.

We've seen security trainings in general, just very point in time and not very successful in doing that. That's number one. We feel that as data privacy becomes a board level conversation point. Currently, when you look into SOC 2 Type 2 or SOC in general, it's a self-attestation. This is what we do. You could potentially, even in your SOC 2 report say that – I'm going to use someone else's example here. I went for a coffee and I'm in my organization safe. Trust me that actually does work for audit firms, if you can convince them. So how do you measure? How are you keeping track of how things are happening and monitoring in real time where your risks are within these – as someone said dark web of your tech stack, which is third party SaaS apps. We want to bring light, we want to have an audit trail, we want to be able to auto remediate some of these risks in these third-party SaaS apps, and really reduce the cost of tech ownership for many organizations who are having to build their own platforms, just because they think that these platforms are not safe. We try to make the box and the Google Drives and the Slack, So the world is safer to use for regulated industries, essentially.

[00:21:56] JM: Let's talk a little bit more about the architecture. So you mentioned us a lot of Lambda based stuff. Can you just give me the rundown of the ingest pipeline and what some of those pieces of a cloud infrastructure look like?

[00:22:11] YA: Yeah. Essentially, the way that we have it set up is, there is an NLP process, which is essentially with the policy. We have a policy engine, which is specific to an organization. The policy engine is your definition of what's sensitive for you, anything which you've set up in terms of defining your kind of rule set. That rule set is an input into your, in this case, a Python process, let's say. On the other side, the data input, which is coming in from your abstraction layer at the top, no matter where it's coming from, is the other input. Those two components essentially kick off this precompiled Docker process, which is essentially housed into a Lambda spawn, which happens. It's essentially running your NLP, filtering on the policies, which has been defined and essentially spitting out what was found in the document.

Then there is another process to be able to kind of keep track for each document what the sensitive data find was. In other processes, basically, keeping track for contextually risk awareness. We want to keep track of how those events actually happened. Which user sent it? Which environment was this sent in? Which people have access to it? Who the receiver of that was or is it a public setting and things like that? Those two things combined, then goes to another process to be able to essentially say, "Okay. It's a risk, or not a risk. Are we going to be able to now decide what remediations we want to take based on that, based on the policies which have been set up?"

It's a two, three, four-point process in terms of then coming up with the remediation actions that go back to the domain platform. Okay, remove the file, or replace the file with a redacted version, so on and so forth. Those remediated actions then can become very specific to individual platform themselves.

[00:24:10] JM: Gotcha. Is there like any interesting scalability questions around how you adapt to spikes in data production from different clients across your own infrastructure?

[00:24:24] YA: Yeah, all the time. I mean, we are using a queuing mechanism, but that queuing mechanism, out of the box, if you use an AWS one or even a Google one, those are not necessarily ideal. In terms of creating rules of prioritizing different queueing order is always a challenge. An event coming in from an internal file share happening on Google Drive versus an email, for example, should take a different way because that has a high risk. That's something which we are constantly looking to optimize.

Then because if there is a peak, at some point, there is some delay which needs to go into the into the queue, which could be SQS in this example, which is actually what we use. How do you optimize that queue? Because if the high-risk event happened, and that got delayed, because something happening in the queue, that could be an issue. SQS so far has been okay for that. We already see some obstacles coming in how SQS queue optimizes a lot of things, or how much we can impact the queue itself. We have considered building our own queuing mechanism, which kind of takes a page book from SQS, and has their own flavor to it. We just have been so overwhelmed that that's one of the lists of things to do for us in medium term.

[00:25:42] JM: I'd like to know more about how a company reacts to finding a problematic data loss event. If there's some situation where the losses detected, do they typically make an organizational change, or a technology change or just handle the event as a one off? What's the response to that?

[00:26:08] YA: The most common things we see is, the surfacing of your risky employees or risky users within your organization. Very quickly, within the first few weeks of going live, typically what we see is like 5% to 6% of your users are responsible for 80%, 90% of your data loss risks, to be honest. That really is like a wake-up moment for most like CISOs, or most directive engineers or even compliance folks, in terms of talking to those people and reducing the risk. Typically, it's been a people change, is most common what we see. Usually, these organizations are committed to these platforms so they don't make that change. We haven't seen that yet. But process change has been something which is based on how different parts of organizations are working together has been another outcome of that.

One example on that real quick has been, for example, we have customers who had to handle with the prior solution, data loss events manually, end of day, they'll have to go in and quarantine the events, and everything was beholden on two persons, this triage team in this large organization of – it's a publicly traded company. But once we kind of came on board, they basically pushed down the owners of managing and remediating these alerts to the employees themselves, where the machine will redact and the users who need to see the information can unredact or auto be able to see things they have access to see.

What ended up happening was the data governance, data privacy owners suddenly stopped becoming a job of a technology team, but it ended up becoming an organization-wide kind of crowdsource kind of problem and a solution to be honest, which was an eye-opening moment for us in terms of – because what we see right now in the market is security teams cannot hire fast enough. They don't have enough people to manage a lot of these alerts. What we see is the future in how organization can keep themselves safe, everyone needs to be involved.

[00:28:09] JM: What's the process for integrating with the output of Polymer? If I want to have some kind of automated triaging system to respond to events in Polymer, what's the typical way of hooking into it?

[00:28:28] YA: You're typically a user of any of the SaaS apps what we connect into, we're constantly adding new integrations. You can come to our website, polymerhq.io or give us a ring. Literally, it takes minutes to install. We're doing just like any other app install of your environment. That's typically the easiest way to get started. It takes literally less than a minute. Once you kind of have tried out the system, and you want to kind of go into an enterprise grade, where you're dropping single tenancy or on-prem install, then typically, that's a conversation to have with us. But usually, that process takes about an hour to get set up. But to try the product out, it's literally a minute or less on any given platform.

[00:29:08] JM: What are the biggest engineering challenges you've had when building Polymer?

[00:29:15] YA: The non-uniformity of the API's of the third-party SaaS apps themselves. Some are stellar, I would say. Some companies who have amazing documentation, amazing API's. While others are extremely sloppy. We see bugs in them all the time. That is always a X factor in terms of time to market for any new integration for us. We have been approached by many smaller SaaS providers, but just the maturity level and how their SDK is setup just causes a lot of angst on our side. That's been the biggest one in terms of pushing, growing the product out in the marketplace.

[00:29:55] JM: Have there been any interesting architectural decisions that you've made in recent memory?

[00:30:03] YA: Yeah. I mean, it is always from a design challenge. One of the things we have to be careful of, it's easy enough for us to bring all the data on our side, on our hosted environment and work with it there. The unfortunate limitation for us as we working with regulated industries, who might not be trusting us as a vendor also, right? Our limitation is that, we can only somehow process metadata. Even if the client data needs to be processed, we need to allow the clients to rip the data using their own KMS or their own keys.

That's been kind of the problem where, what we tried to do then is, initially, we didn't have a KMS capability. Then we had to quickly go ahead and build that, where both the data in motion coming in from the SaaS platform, back into our abstraction layer in the backend with the Lambda processes. But also, any interim storage of the data happening within S3 or other systems of record for transient data or copies of data while we render something out. For example, all that data, we had to basically make sure was encrypted also. Then we had to quickly push out a way to make everything Dockerized, everything being able to be encrypted with a KMS key, or a key chain at the client side. That was a big lift for us, which we didn't anticipate when we went into this business.

[00:31:23] JM: Can you give a little bit more context on why a selection of a key management system would be so important?

[00:31:32] YA: I mean, it's my lifeblood, like my Google Drive environment, my Slack environment, my email environment is literally my business. If I'm adding a vendor who needs to process these things by factor of X, I'm expanding the risk profile of my company. SOC, and all those things are great, which are like, we think that those are awesome, but they're only self-assessments. We are assuming that no one – as a customer, you don't need to trust us in terms of how we handle the data. Even if we are breached, your data is secure and that's the overriding principle in how we delivered the solution. KMS is essentially, is a page book we actually took out of snowflake, their go-to market when I was doing my consulting thing. That was one of the things they implemented to be able to sell databases to their customers. I saw that work for large insurance companies, large financial services. That's why we adopted it, kind of from that experience.

[00:32:34] JM: Does Polymer help companies reach a certain level of systematic compliance that they might need to achieve? Does it help with any particular compliance regulations or is it more of just a general security protocol?

[00:32:51] YA: HIPAA, definitely. We are HIPAA in a box, I would like to say. In terms of many of the platform you hook in, you can apply operational HIPAA as well as data storage HIPAA, like true HIPAA governance. Then the second one is SOC 2 type 2 data control component of your type two assessment. SOC 2 type 2 requires you to prove that you have controls, which are working and wishy washy is kind of how many audit firms are allowing companies to go about it. Our product basically allows a demonstratable auditing, as well as remediation to ensure the audit data controls within your organization. Then thirdly, kind of GLBR field or banking regulations, those are also covered but they are less frequent as use cases for us. GDPR, obviously, is there but GDPR gets a little fuzzy, to be honest.

[00:33:47] JM: Do people even care about GDPR anymore?

[00:33:49] YA: Only on paper, to be honest.

[00:33:53] JM: What are the areas that you're working on right now in the platform?

[00:33:57] YA: We are expanding the integration base, for sure. That's number one. That's always going on. One of the things we have been getting a lot of requests is, from an insider threat perspective, especially for financial services firms. If someone is downloading a lot of financial models today, or they're emailing them to someone external, and they have not opened a financial model in the last 30 days, does that constitute to be a suspicious activity? Does the compliance officer or your CISO needs to at least take a look at that? Those kinds of telemetric data around user behavior are something which is becoming kind of an interesting angle for us. We're assessing what's in the document. Is it a financial model? Is it a ProSAP? Is it a customer file, for example?

Then be able to triangulate that with user actions. Who's doing what with that data? Especially as compared to what they were doing in the past X weeks, or X days or X hours or so. That is one area which we are really kind of bullish on in the near term and we are adding a lot of functionality around highlighting some of the suspicious activities and triggering, right now alerts. But then, very quickly after that, auto remediation, switching have access to users automatically if something suspicious is going on. Because we have also seen, with some of the breaches happening, a lot of times, the hackers would assume the IM persona of employee, and then go inside and start like taking a lot of files out, encrypting a lot of files out within these platforms. Because SaaS is becoming a very major source of your record of truth, we are treating this like any other environment historically, like a database or S3 folders, for example. We want to bring the same amount of regimented security within this environment.

[00:35:48] JM: What's been the most effective sales process for distributing Polymer? Is it mostly inbound or do you do advertising? What's your method of reaching potential people?

[00:36:02] YA: No to advertisement. We have lot of like premium users actually. Just the ease of use, to be honest, has allowed a bunch of like mid-sized businesses use us. They never talk to us. Some of them convert. That's been great. We just have been, to be honest, we haven't had a salesperson until like literally last week. We will be starting to kind of make more concentrated outgoing effort. But one observation I've had like two years ago, when we were talking about SaaS security, SaaS Data Governance, people will laugh me out. Now, that has changed a little bit where we are seeing a lot more in the press and people are looking at data governance, data privacy as the other side of the coin of data security. I feel that the market has shifted people's view on SaaS has shifted also. So it's becoming a much more easier conversation just out in the cold also. But our self-serve go-to market motion is something we are very proud of. And we are looking to expand on that very aggressively in the next few months.

[00:37:03] JM: We didn't talk much about your graph engine. Can you talk a little bit more about the data structure of the graph database and what you're actually graphing, and how you take advantage of that when you're executing Polymer?

[00:37:19] YA: Yeah, no. That's actually is why the name Polymer initially came about is because the whole company was built on a graph database in the backend. What graph database allows us to do is look at schemas radically differently very easily. For example, within a user, let's say user node, what access levels they have or what entities they have access to. That's an easy picture to make and those entities then can be coming in from, let's say, another node, which is your SaaS platform. All these interrelationships around user permissions, events that are happening over your platforms, and then be able to query that data and make decisions around what actions you want to take. That's been a big plus for us.

We do use kind of Postgres and other databases for some of the more conventional reporting kind of dashboarding and stuff. But graph database does allow us from this user permission management at an entity level. Those entities then are impervious in where they reside. That particular data structure for us has been very helpful in speeding up some of the querying and action taking. We plan on kind of adding more type of entity nodes in there, for example, not just looking at entities as nodes, but also document types of nodes. Is this a financial model? Rather than caring about what's within the financial model, we can just – for some financial services companies, they're not so much concerned about the PHI. They are concerned about document tags themselves. That's a different layer of kind of document tagging, rather than an ER model, which we are currently employing. Graph database is highly scalable, and highly customizable for different use cases with the same data within it, which has been a big plus.

[00:39:10] JM: What does the read and write look like for that graph database? What's the backing database for the graph?

[00:39:18] YA: It's Neo4j. We are big users of Neo4j in the past, and we kind of carried it over Polymer.

[00:39:25] JM: Okay, cool. So I guess a read and write is just same as it would be for Neo4j. Have you built anything on top of it? Have you augmented the database with any particular, I guess, kind of homegrown algorithms for how you're crawling it and how you're assessing – like you said, you have to do a lot of nearness calculations, right?

[00:39:45] YA: Yeah. We have done it that way. What we did was essentially store some of the more event type data within a PHP database, because just because the performance – sorry, in the Postgres just for performance reasons, because all these events coming in, we want to store them very quickly. Storing and changing the structure in the graph database could be slow, especially as it grows. We didn't want any latency in that. Then based on that record, querying the database or pushing the stuff after the fact, as an insert later, is something which has worked for us because some of the stuff we are looking at graph databases is not for real time actions. So combination of Postgres, along with graph database for us has worked well, where some of these scalability challenges which you run into graph database we want to avoid, especially as it grows over time.

[00:40:36] JM: Well, as we begin to wrap up, is there anything interesting on the engineering side of things that we haven't covered that you'd like to highlight?

[00:40:44] YA: It's constantly evolving field. We're constantly experimenting with new things. We are hiring, so if someone's interested, we'd love to chat.

[00:40:50] JM: Well, thank you so much for coming to the show, Yasir. It's been a real pleasure.

[00:40:52] YA: Thanks, Jeff. Really appreciate it.

[END]