EPISODE 1171

[INTRODUCTION]

[00:00:00] JM: Data leaks can cause privacy violations and other cloud security vulnerabilities. Visibility and control of cloud resources can help secure data and ensure compliance and governance. Open Raven is a system for discovering and classifying sensitive data in a public cloud and assuring compliance and governance. Dave Cole is a founder of Open Raven and he joins the show to talk through data protection and what he has built with Open Raven.

If you want to support Software Engineering Daily in a greater capacity, go to softwaredaily.com and become a paid subscriber. You can get all of our episodes without advertisements.

[INTERVIEW]

[00:00:40] JM: Dave, welcome to the show.

[00:00:42] DC: Thanks for having me.

[00:00:44] JM: So you work at Open Raven. Explain what Open Raven does.

[00:00:47] DC: Yeah. Open Raven is aimed at solving the problem with big data security. So when we started the company early last year we looked out and we thought with the move towards protecting the identity, security's been through a series of phases. Back in the olden days it focused on protecting the network. Then people moved well beyond the network. Mobile came in. Everything else we moved to protecting endpoints. Now there's a very rightful focus on protecting identity and Octa seems to be the de facto leader of that. We looked out towards the future, my co-founder, Mark and myself, and said, “Isn't it logical that the next big wave is that we start protecting these enormous mounds of data which seem to be leaking everywhere?” And we looked at the space and there was a lot of older companies that were

focused on on-prem data stores and doing things in ways that made sense we felt like five, ten years ago focused on email endpoint, office files and so forth. And we looked at the new school companies and they were all very compliance-focused and very privacy-focused, and we completely get that. It's a non-discretionary purchase and so forth to do things for GDPR and CCPA. But in our heart of hearts we looked at it and said, “We think compliance is really the byproduct of doing things the right way,” which means hardcore data protection, focused on where we think the biggest issue is going to be, where the most data is going to be, which is in the cloud. So Open Raven is aimed at discovering, making it super easy to go out and discover all the data you have in the cloud. Making it easy to classify it at scale, wrap policy around that. You can think of it as sort of rules for how data should be and then taking the actions to make sure that those rules are followed.

[00:02:36] JM: Give me an example of some policies that you might administer.

[00:02:39] DC: Yeah. So easy one. Anytime when you find developer secrets out there in S3 and it isn't “properly protected”, meaning access permissions aren't set up appropriately on the bucket or so on and so forth. Give me an alert. And the contrast to that would be kind of a more classic security alert of, “Oh my god! You have an s3 bucket open.” Only to find out that it’s marketing content. So we sort of marry together what you might know and love or not with AWS Macie or like a Google DLP alongside the rules you might have inside a cloud security, posture management product like a Cloud Custodian or a Palo Alto Prisma and so forth.

Other examples include things like we can only have this type of data. Maybe it's personal data. Maybe it's source code, you name it, in this region or in this location. It should never be anywhere else. So let's say you know you're a tele-health organization and you have an e-commerce front. The HIPAA-related data needs to stay over here. It can never go into the e-commerce side of it. There're tons of examples, but a lot of them focus on either finding data that you're very concerned about like developer secrets or personal data keeping it in the right location, keeping it configured properly and so forth. Those are some of the top of mind things.

[00:04:02] JM: Tell me about some of the common anti-patterns that you see in open cloud security.

[00:04:07] DC: Yeah. What we're seeing a fair bit is just sort of baseline issues with not knowing what's where. So on our side, and what prompted all of this when Mark and I were looking at what problems to take on last spring, was really the question of where's my data kept coming up. We originally thought we might take on asset, more of an asset management type problem. Once we got into it, we realized everyone said, “Well, if you could tell me where my assets are, I could then pinpoint my data.”

So foundationally, like the biggest issue we see is that it really starts with I don't want to call it poor data governance, but simply knowing what data is out there. S3 has become sort of the junk closet of the public cloud and it’s got a bit of everything in it once you get its scale. You bring that into a multi-cloud setting where you've got not only multiple teams in a single cloud but you've got multiple teams across cloud and the problem quickly becomes just a sprawl containment hurting cats. So the first one we see is just how in God's name do you keep your arms over this? There's certainly things like AWS Macie out there, but we've heard pretty consistently it's simply too expensive. And to-date, it hasn't worked beyond S3 and so forth. So you've got a very real problem of things like developer secrets, whether it's source code or you've got things like shared environments with a partner where you have sensitive data sitting inside a shared environment that everybody's responsible for, but nobody's responsible for at the same time. Or there're things in the contract that were laid out cleanly as to how it should be handled, but there's no way of enforcing it. So it's really a large number of things and a lot of them come down to too many people touching data, vast amounts of data, access permissions being kind of ill-attended to and things moving incredibly fast and the entire environment being a little too prone. It’s very easy to have an accident. And the accidents are what's behind the daily parade of data exposure, data leaks, data breaches.

[00:06:25] JM: So how do you actually program those kinds of issues away? What do you do or what's the surface area of the product that you expose to the developer to allow them to program away those kinds of problems?

[00:06:40] DC: Yeah. So we're more focused at the infrastructure team. So much more of like a DevOps SRE or like a SecOps team than the developer themselves. So I'll answer the question from that perspective and I'll sprinkle in the areas where someone who's maybe more of a classic developer would plug in. But the way that we tackle this is first we start out with the assumption that you don't necessarily know what's in your environment. So on one side we'll use the native APIs. We'll use things like AWS Config, we'll use Cloudtrail. We will call the APIs that are available.

For things like generic compute sitting out there, so EC2. We have a lambda-based analysis model where we will fingerprint data stores sitting out on EC2. So non-native storage requires a little different approach. We built a ML-based classifier that executes over lambda. So no agents, no scanners. And that brings it all back into what we render in a 3D map. So the 3D map itself, if anyone's interested, is built in Babylon 3D, and that makes it easy to visualize. It's also pumped into a database with a firehose API and so forth if you just want the asset inventory.

From there, the classification is done over the same lambda model and the classification has to be completely configurable. So what you would do is you would go in and say great. We see Open Raven that you have a template for something like personal data or PII. PII means something wildly different and not wildly different, but considerably different in every organization. If you're a wearable company, it's probably the serial number on your Fitbit clone or whatever it is you're wearing. If you're an auto manufacturer, it's the VIN, and so on and so forth.

So you go in and you can modify that to be really whatever is necessary to your business. That basically gets configured as a job, as a classification job that you can run on a schedule or you can do it on an on-demand basis as things change inside the environment. And then that basically builds you really rich context on not just the assets that are out there, which was that first step, but also the data that's sitting out there as well. And the place where those two come

together and where you act upon that context is in the policy. And for that we use Open Policy Agent.

So to come to your developer question or an SRE or even a security person who are increasingly developers themselves, they can write Open Policy Agent rules that stitch together logic about how data should be handled by the resources that are out there. And since this policy is code, they can do that in a pre-production and they can also do it in a production setting in order to make sure. So, clearly, if it's in pre-production, they can see pre-flight. You know what they're pushing out. Is it going to be a violation of data protection standards, whether it's security or privacy? And if it's in production, when someone's, let's say, put an API key, that shouldn't be out there inside an S3 bucket that's exposed. You can catch it and catch it in production before someone else finds it just to use an example.

And the responses are configurable and lots of security companies will talk about, “Hey,” and then you can change configuration or block things or so forth. The reality is most people just want an alert at that point and they either want it written into Splunk. They want it written into – They want an SNS message, you name it, and they can take action from there.

[00:10:27] JM: You mentioned Open Policy Agent. Could you detail more what OPA is?

[00:10:31] DC: Yeah. So open policy agent, it's been out there for some time. So it's certainly not new. We discovered it when we were looking for a great way to bring in a policy agent. We knew he needed it in order to really for kind of rules logic and to have a rules engine inside the product. And it's this really vibrant open source project that's created a strong scalable rules engine with a pretty straightforward language to write policy rules. And there's a whole bunch of folks using it today. So it's got a nice vibrant community. And really, it's part of the incredible thing of building products right now. Like as little as five, ten you know years ago, you simply wouldn't have been able to find something like OPA out there to give you a really robust rules engine to build policy rules and so forth. You would have had to license it or build it yourself and it wouldn't have been nearly as good.

So there're folks using OPA for all sorts of things. We were comparing notes with a company yesterday, a big internet company that's using it for a whole bunch of operational tasks and just automating, checking on resources. They were using it to monitor things like SSH connections and so forth and apply logic. So it can be used for any of a number of things.

[00:11:53] JM: So walk me through the process of setting up Open Raven and what I would do to actually get my assets secured.

[00:12:03] DC: Yeah. So you've got a choice. Historically, the way that we designed Open Raven is we designed it to be deployed into a customer's own VPC, and that's how the current community edition is designed. You deploy it. It's effectively a Kubernetes cluster. We remotely updated over Helm charts, but effectively it is a single tenant system that sits inside the customer's own VPC that we update remotely. Telemetry comes back, health signals, licensing, all that happy stuff. But that's the deployment. It can be done with Terraform. It can be done with Cloud Formation. It is AWS only today. So that's the one caveat. We have ambitions to go well beyond AWS, but the platforms is – I think anyone listening to this would understand are different enough to where if you want to do a really good job. You've got to focus in the beginning.

So from there, once you've actually done the deployment, which typically takes 15 minutes tops in order to set it up, you go about setting up discovery configuration. So how you want it to access your environment? So you can go small and add in a cross account role and maybe see a small environment, or you can go big, or anything in between. And going big might be you have hundreds of AWS accounts inside an organization and you add the organization itself, which of course adding it in by the org means it's somewhat future-proofed. Anything that goes in the org in the future is automatically discovered much more powerful, but most folks like to start with a little smaller slice of things. Get to know you. See how it works before they do that. So that's the logical next step.

From there, you configure data classification. You modify it to your taste for anything that you're specifically looking for inside your environment or you can just use a template. Same

thing with policies. You can pick an off-the-shelf policy like an AWS CIS benchmark if you want to do just a quick kind of data risk posture assessment. Or you can go ahead and say, “I know what I want. There're these specific things that I'm looking for. And that's what I want to build my policy, my logic for.” And after that it pretty much just runs on the schedule that you provided. If you're an organization who wants it to run in a real-time capacity, it can run in a near real-time capacity. If you want to do just periodic assessments, you can do that too. It's really up to the preference of the organization.

[00:14:35] JM: So how often is – Is there some kind of runtime model for analyzing my infrastructure for issues for security holes? Is there some ongoing oversight that runs like a cron job?

[00:14:53] DC: Yeah. So first off we, get eventing through CloudTrail. So that's part of it. But let's say that you want to continually classify data, either you want to do it periodically or you want to do it on changes in the environment, meaning material changes have been made to this S3 bucket or any S3 bucket. So in that scenario, as before, the analysis is lambda-based so that you don't have to deploy an agent or a scanner and it kind of solves some of the budget issues that folks have had with things like Macie in the past. So there, if you want to set it up in something in real time, you'd be looking at events. Events would be coming through us through CloudTrail, which is nice. But if you wanted to go ahead and do continuous data classification on change, you can do that as well. We'd be doing that over lambda, and off you go. It's quite configurable. We have folks who are more interested in doing routine assessments and you can set it up so it's not doing kind of real-time monitoring, or you can set it up in a full kind of pseudo real-time data protection monitoring basis. It's really up to the needs of the organization.

And what we're finding is at this point we're doing a fair bit of hand holding and making sure that we get it right for folks. There's a lot of configuration with classification. We do some things in it such as the ability to call out to an API to make sure we got the classification right. So for an example, with a payment card, you can do something called a loan check to make sure that it's a legitimate credit card. Other folks have APIs for things like game codes or

verifying a vehicle identification number. We're doing a lot of the – Helping folks set those up now and creating rules for policies if they're not familiar with Open and so forth. So it's an area where a product can do whatever you want. The platform can do whatever we want. We're helping folks kind of get the most out of it now. In the future, we'll probably take off some of the effort there, but right now it's probably a little more developer friendly than not.

[00:16:56] JM: What's been the hardest part of building Open Raven from your point of view?

[00:17:02] DC: COVID's been a massive disruption. The technical challenge is getting the data model right. We've gone back and forth on some – We started with OrientDB. We've used Elastic. We're finding that we'll probably change the backend one more time, so the data model itself. And this is true with so many platforms, is the data model is a challenge. I'll say like that's probably the biggest technology challenge. Other than that, it's been – Most things have been relatively straightforward. 3D mapping has been surprisingly easy. I think that's just due to some of the good help we got. Our HQ is here in Los Angeles and there's amazing gaming talent here in LA. So that's the one that's actually gone a little faster than probably what we anticipated. That and some of the data classification work. We got lucky in a few instances.

The most challenging thing has just been plugging away during this year when no one can see each other. And teams move as fast as they trust each other. Trust in my estimation is predominantly built through kind of rich, high-fidelity interactions. And those are really hard to come by this year for means of safety. So I'd say like the technology challenges, the biggest one has been getting the data model right, getting the data backend right. The biggest challenge overall is moving fast as you would imagine a startup to move onboarding new people and doing it without being able to see each other well with a backdrop of kind of apocalyptic conditions. So startup world is an extreme sport. This is super extreme.

[00:18:47] JM: How were your past experiences in various executive roles, how are those formative and how you decided to build Open Raven?

[00:18:57] DC: It's a great question. A couple things. So one, remote work itself. CrowdStrike was remote first. So I had a lot of good experience with what works and doesn't work for remote work at CrowdStrike that informed our model as well as my time at Tenable, which had an office culture and then moved to remote. So that was heavily influential, as was my co-founders experience over at Source Clear before this. It was a company he was CEO at. So that was a big thing, is we looked at it and said we're absolutely remote first, but we're going to get together in-person and here's how we'll do it. And we only hire West Coast, unless someone has really proven that they have the maturity to work in another time zone and work async style.

Another piece is we looked at this and said the model of the future we believe is open core, and we think that's where it's going. We've seen companies and partnered with companies that we think got the model right and they'd earned our respect. So that was a big factor, is we wanted to have a product-led company with an open core design, with an open core model. The product-led has come from being in organizations that were very sales and marketing driven and seeing how hard it is for the product to get the right level of investment and how hard it is to make good on the customer promises if you're not leading with technology first. So that kind of experience, and really kind of some of the – what we call the bad behaviors we're seeing in the security market across the board of over-marketing, overselling and the product not delivering really informed the fact that we need to be tech-heavy. So we've been only recently adding in a handful of sales and marketing folks in order to complement a very product-driven strategy.

[00:20:54] JM: Tell me more about the usage side of things. What does the user use for Open Raven? What do they do?

[00:21:01] DC: Yeah, so it depends on who you are. But our senses is this and what we're seeing from our early design partners is they use it in the way that you might use something like a classic monitoring product, like a Datadog or a visualization product as well. So we had one customer said, “Oh, wow!” And this wasn't our intent. They're like, “Oh, instead of going into AWS console now, I go into Open Raven because I can get this cross-regional view of

everything and I can see it there in a way that I couldn't before and I can start to dissect it and kind of pull it apart." That's kind of one use case.

What we're seeing there is people right now are in the stage of configuring classification jobs, going through the results, going through what's back, hunting through it. There's a fair amount of use cases here that are not unlike what folks use Splunk for in security and for kind of ops analysis where they can get kind of a pile of data and pick through it. So there's some interactive usage. There's a fair number of people who just want you to dump the data as well, hence the firehose API, into something like Splunk or into something that can to Splunk and then just pick through it there and see it next to the rest of their security data. So it depends on the persona, it depends upon the person.

[00:22:33] JM: How do you test Open Raven?

[00:22:35] DC: Yeah. So we use Selenium. We're a Sauce Labs company, so we use Selenium for kind of frontend testing and so forth. We do security testing as well. So we use White Source. So that's what we use for our security testing, pure programming. There's kind of no substitute for looking at each other's codes and dialogue there. In addition, we also still do some outsourced testing. So we have a partner that I've used at CrowdStrike, at Tenable and we use them at Open Raven as well where they build some of our test automation, but they also get hands-on keyboard and do some actual tests up against the UI as well. So it's a combination of things. Everything from test automation that we build with Selenium or other tools. We do some outsource, kind of hands-on keyboard testing and including security testing. We also pay for security testing too. For what we're doing, we think it's kind of unacceptable not to have a third-party penetration test. So that's part of what we do too.

[00:23:42] JM: Can you tell me more about how Open Raven discovers my assets? Because I could have complete infrastructure sprawl.

[00:23:50] DC: Yeah. Yeah. Matter of fact there's a number of data discovery products that don't start with that that assume that you know where your data stores are. What we've found

is that most people don't. So we do a number of things. So first off, we'll call the AWS API. So we will use conventional AWS APIs. We'll use AWS Config and so forth. There's no need to do anything crazy and start blasting packets across VPCs and so forth. We use the APIs that are there in order when we can. And then for places where we want to identify the rest of the resources where it's sitting on EC2, that's where we do lambda-based analysis. And effectively it's our Dmap classifier. So there we're not looking for every type of resource. We don't consider ourselves kind of a generalized today cloud resource asset solution, but what we do is we focus on identifying, pinpointing all the data stores that are out there. And if you're wondering what that looks like, there's a list of those up on our website under product. What we discover. And that is a straightforward ML classifier that is – I believe today it detects 80, somewhere between 80 and 100 different data stores. It's the top data stores as listed by DBEngine minus a few things that didn't make sense like MS access.

And with that, what it does is over lambda we're effectively running a port scan. We're running the classifier. It's built to also work on the file system. But having said that, today it's just easier to run it as more of a port scanner and there it uses the classifier in order to fingerprint the data stores that are out there. So we're doing a combination of what we think are pretty common sense things using the existing APIs when we can and then we're using this lambda-based analysis model with – In this instance, a straightforward ML classifier in order to find the data stores. And we pull that all back. And again, you can access it programmatically through the firehose API. We've been using GraphQL as well for folks who don't want the firehose but want one-off request, or you can look at it in the pretty 3D map. So you can you can go fluffy and look at the really pretty version or you can get as down into the bits and bytes as you want to.

[00:26:20] JM: Can you tell me more about building that machine learning model?

[00:26:24] DC: Yeah. It's interesting. Honestly, it's pretty straightforward. So we looked at it and said, today, would you really – In the olden days what we do with this is we do a combination of kind of port analysis and banner grabbing and that sort of thing. With it we looked at it and said, “Let's future proof this a bit.” And there's enough consistency across the

different types of data stores to where they have enough of a personality through port interrogation or by looking at the file system that it's worth an ML classifier.

Having said that, so what we did is we stood up a whole bunch of data stores, and this is where we used an outsourced partner as well, stood them up in Fargate. Did feature extraction both over the file system over the ports and iteratively tested and made sure that it was accurate and then also built in a feedback loop so folks could tell us when we were off, when we misidentified something.

As far as ML classifiers go, there's no black magic here. We did things that I think were pretty sensible that were better than just doing an old school kind of service identification, service fingerprinting algorithm. We think it's more future proof. It's got a nice feedback loop. But having built or been on teams and had teams that built ML classifiers before for things like malware detection and so forth, this is pretty straightforward to the point to where we posted a white paper on it for anyone who's interested. So you can go in and get really detailed information on it.

We're starting to do more sophisticated things now with data classification and machine learning, but even there, we found that things like data adjacency can work vastly better than machine learning at a lower cost and be easier to maintain at times than ML. So the ML classifier, it works. We're proud of it. It's great. But having said that, I think people overplay the impact of ML today and it's been applied to things where at times a simple regex are just fine. Or in our case we're finding that data adjacency, meaning if I see a 16-digit number next to a 3-digit number next to a 4-digit number with a slash in the middle, like that can work way better at less cost than trying to do an ML classifier. So we're using it, but you'll never see openraven.ai be a big thing for us. We think AI has been – The application in our space is interesting, but it's kind of far from the linchpin of the technology.

[00:29:06] JM: Do you do anything around threat detection, around detecting if a an asset has actually been compromised?

[00:29:14] DC: Yeah, it's an interesting question. I've done a lot of that in my day both at Symantec and at CrowdStrike, a little bit at Tenable. We don't today. That doesn't mean that we won't. For example, one of the things we can do is we can very easily give you a list of every file that's out in S3 in a speedy fashion. What we could do on top of that is we could hash those files. At that point we could run those hashes or someone could do it themselves with like a YARA scan up against the list of known indicators of compromise and tell you pretty quickly whether you had malware there or something matched a known attack tool and so forth.

So it's certainly within reach. It just hasn't been a focus for us to date. But that's no reason we wouldn't go after that and provide it, but it hasn't been a core focus. We've been much more focused on the data leakage front. But my gut tells me we'll end up doing that at some point and even providing the ability just to signal people on that automatically. But I think there's a whole bunch of folks who have ambitions in that space, in the threat detection space in the cloud in particular. The piece that we felt was missing was the data inventory and classification and data location at scale and the ability to quickly apply rules to that and build custom rules around it. But yeah, we may end up going back into that space if there's enough customer demand.

[00:30:45] JM: How big is your engineering team?

[00:30:48] DC: You know what? Today the company's 25 people, and that's probably as big as we'll be for a little bit. We've hit the scale where we can support current customers and the next round of customers as well. We've got plenty of capabilities engineering and product. I would say product in general, we have designers as well who are 100% on the product team even, our creative director who's at times mostly dedicated to product. We're about 70% product of those 25 folks. So depending on how you count it, it's 16, 17 people.

[00:31:23] JM: And how are those teams arranged?

[00:31:25] DC: We use feature teams. So even at our small size, we have feature teams. We feel like that promotes the right level of accountability and control and candidly just makes the team happier. We weren't doing that before and we stopped and retooled this summer after we put out our community release of our product and moved to a feature team model. We'd always planned on doing it. This is a little earlier than what we anticipated and are already yielding really good results with a feature team approach. So that's classic two-week sprints.

[00:32:03] JM: Are there any kinds of assets, cloud assets that are particularly hard to discover like things that just may be in the weeds somehow and you can't scan for them?

[00:32:17] DC: Ooh! At this point we haven't seen anything like that. We've had a few surprises. So I'll give you an example of a surprise and something we found that was unusual, not exactly the answer to your question. But the one thing that we have found is some incredibly old kind of first generation EC2 out there. EC2 instance from 2013 is something we did not expect in any way that was exotic, and that's just sort of the game of working with larger organizations. They're going to have that out there.

Today nothing's been exceedingly hard. Matter of fact we just added a whole bunch more resources. So initially, we were very, very focused on fingerprinting data stores and predominantly on data service APIs and so forth. We've just expanded out I think to a whole bunch more resources in order to provide a richer set of cloud assets to write policies against. But in fairness, I don't think we're deep enough in that to say like we've learned the personalities of those and if we're missing anything. It doesn't feel like we are at this point, but it's early innings.

[00:33:35] JM: You work with each of the major cloud providers. Is there any nuance between Google and Azure and Amazon that's worth pointing out?

[00:33:43] DC: Today we only work with Amazon to be clear. So our chief architects from Azure, and we know enough to be dangerous about GCP. But having said that, plenty of nuance. We've been exploring at least the differences in the security models on the podcast

that I run security voices. But having said that, at Open Raven, we've focused exclusively on AWS in the beginning. So the platform is intended to accommodate any cloud service provider. But today, in order to make progress and go deep, we decided to go deep in AWS. So up for us next after – Later after S3 is RDS. So we're going after structured data next and Snowflake, and up after that is GCP. But it's around the corner. We're not there today.

[00:34:34] JM: There's this term drift, configuration drift or infrastructure drift. What causes drift?

[00:34:42] DC: Oh boy! It's a good question. I think there's multiple causes, but it's speed and pace and the number of people who have their hands on the dials. The number of accounts, the number of people, the crazy rate of change inside projects and versioning. I think there's very few organizations who have a really disciplined set of controls that would prevent drift primarily due to the fact that they prize things like expediency and getting clean burn downs in their sprints, getting functionality out more than a prize governance, which may or may not be the right business decision. Certainly as a young company, you have to make tradeoffs for speed. So config drift, I think it can come from a number of places. But having said that, it's far too easy to do, and we've experienced it ourselves.

[00:35:40] JM: So as you said, Open Raven has an open core model. Are there a lot of users of the platform that are contributing back to the open source project?

[00:35:50] DC: No. There isn't today. A matter of fact we looked at what we'd done from an open core perspective this past summer and decided it simply wasn't good enough. And we've recently brought on David Lester who comes to us from Apple where he ran the FoundationDB project in order to get our open source strategy, our open source activities in order. It's just that classic thing of small company startup. Unless you get someone focused on it, unless you get someone dedicated to it, it just doesn't happen in the right way. So we're in the process of rebooting our approach to open source. We have some things out there today. Some folks have used those. They're primarily focused on data ingest. I think the closest analogy to what it would be like is maybe like the cartography project, what we have out there today. But we

don't think it's nearly enough in order to foster the kind of community and engagement that we think we should have. So that one's in the process of being rebooted. So look for more on that actually when we announce our paid product in November and even a bit more early next year.

[00:37:05] JM: When you talk to the kinds of large enterprises that you do talk to, are there any other security issues that they are battling that we haven't really explored?

[00:37:17] DC: Yeah. Well, this year has been unusual to say the least. I think a lot of what they've been dealing with is just grappling with remote work. And it's everything from accelerating things like zero trust project and multi-factor authentication and shutting off old VPNs to bolstering VPN capacity because they just can't swap it out fast enough and it's too risky. So the bulk of this stuff that they're dealing with this year has been quite a bit different than before. Well, it's been acceleration.

So on the cloud side, we've seen people focusing on just getting things deployed, getting things out there, and we are seeing there's still a lot of interest just in getting baseline configuration right, cloud security, posture management. We certainly see a lot of that and focus on cloud asset. Beyond that, there's still a ton of interest in manage, detect and respond. So security operations centers are moving up the maturity curve. Some folks are just handing it over to others. So you see huge financing rounds being done recently like the one that was done by Arctic Wolf. I think they pulled in 200 million. And I think once a week I'll talk to someone who's in the manage, detect and respond space where people are just saying, We don't want to run our own SOCs." So there's quite a bit of that. Either people are really investing their SOC and running it by themselves and using orchestration products in order to automate a fair amount of security, or they're tapping out and handing it over to someone else who will do it for them. That's just a little bit of what we're hearing. So a lot of zero trust, a lot of blocking and tackling related to remote work. Certainly we're brought in on the data protection side. So data security, data privacy, and then a lot of work either done improving your SOC or just handing it off to someone else to run.

[00:39:12] JM: I think there are a lot of engineers out there who mainly work on application infrastructure and don't really spend much time thinking about security, because somebody else at the company takes care of security. Are there any areas of security that you think would be useful to describe to the average application engineer who doesn't know much about security?

[00:39:34] DC: Oh wow! That's a great question. Probably a better question for my co-founder. And let me think on it for a moment. I think the area that's been getting a lot of attention is the one that my co-founder comes from, which now seems to be popularized by Snyk, which is probably familiar to application engineers actually. So it may not get at the heart of your question, but there's so many potential issues with software supply chain and just vulnerabilities being passed down and then arriving inside your software as components. One component calls another, calls another. That's the primary one that comes to mind. It's probably well-known to many whether or not they're using something like Snyk or the product that Mark brought to bear, which was called SourceClear, which is eventually bought by CA Veracode.
I personally think that's one of the more interesting areas.

So software supply chains an area where there's a lot of interest and a lot of concern and increasingly a number of products and capabilities. Snyk seems to be at the forefront of that and one of the most popular choices. But beyond that, I think a lot of the security issues at this point, at least from an engineering standpoint, are reasonably well-known. Is there stuff in the threat landscape and threat level that's interesting? Sure. In particular, there's stuff that's happening with industrial, in the area of industrial security. That's particularly interesting, but it's probably a little too far afield for application engineering and people – Developers might be interested in security. Certainly that IoT area with an emphasis on industrial security, there's some really interesting things happening there now. But like I said, I think it's pretty far afield for probably most people who are developing applications.

[00:41:31] JM: When you examine one of these data stores, are there any particularities you have to program to have all the coverage you need so that you cover Amazon S3, Red Shift,

MongoDB, Elasticsearch. Is there a lot of work you have to do to cover all of the database, potential databases out there?

[00:41:53] DC: For the baseline classification, for the baseline identification of a data store, like yes, there's work that has to be done. We have to make sure that we can fingerprint the data store itself. Next up for us is the inventory and classification of the data, and that's where it's incredibly specific. S3 is where we're focused at the moment. We'll move on from unstructured data in S3 to RDS. And each one of those is an incremental lift and each one of those particularly going from unstructured data in S3 to going to structured data in RDS is a substantive incremental work. And we expect it'll largely be the same for the different data store types.

But for example, early looking at Snowflake. It isn't that much more effort. So we're pretty fortunate we've got a great architect, a gentleman named Mike Andrews who's built the data inventory and classification, which is the more bespoke piece of it in a way in which we can we can go after those fairly easily and where we can kind of update and move. But none of it's free. It's a fair amount of work. And the more different it is from currently what we're doing with unstructured data, the more work it is, which I think follows pretty logically. So stay tuned, but it's considerable amount of work. But the hardest work as always is getting the initial scaffolding in place, getting the first one done. After that, it feels like it's probably about 30% to 40% of the original effort.

[00:43:25] JM: Tell me about the future of the company. Where do you go from here?

[00:43:29] DC: Yeah. We want to be at the end of this. We want an organization to be able to any point in time know where all their data is, know what type of data is there and make sure that it's protected in the right way that's safe, it's private and that any compliance report, anything that have to do is really just a click away. All of this pain they have, the manual effort is gone, and data leaks, data breaches inside the organizations that use the product are a thing of the past. We want to be that back plane for data protection to where they can ask and answer any question about their data particularly from a safety, from a security, from a privacy

perspective and have it readily answered. That's our goal. So really for us, it's just marching one cloud environment to one cloud service provider to the next, to the next and adding in –Going beyond infrastructure as a service and going into SaaS applications as well so that it wouldn't just be about something in Azure, Google or Microsoft, but it could be something sitting inside a SaaS application like a Slack or anything that might be sitting over in GitHub and so on.

[00:44:39] JM: That seems like a good place to close off. Is there anything else you want to add about Open Raven or security in general?

[00:44:45] DC: No. That's good for my side. Thanks for having me. I appreciate it.

[00:44:47] JM: Absolutely. And I do want to point out to the listeners that they can find you on your podcast, which is Security Voices, I believe?

[00:44:57] DC: Yeah, that's right. It's securityvoices.org.

[00:45:01] JM: Great. And that's stories about security. And I encourage people to check it out.

[00:45:07] DC: Thank you.

[00:45:08] JM: Thank you, Dave.

[00:45:09] DC: I appreciate it

[END]