Synopsis

Join Corey Quinn of Last Week in AWS, and Open Raven Co-founder and CTO, Mark Curphey as they talk about how security officers don't know where data is, don't know what data they have, and don't know how that data is being protected; how companies that manage other people's data have a massive obligation to protect it but few take that seriously; why Mark gave up trying to sell developers on the importance of security; what the OWASP is and the origin story behind it; the increasingly important role security will play in the data economy; Amazon Macie's terrible pricing story; and more.

Transcript

Announcer: Hello, and welcome to *Screaming in the Cloud* with your host, Chief Cloud Economist at the Duckbill Group, Corey Quinn. This weekly show features conversations with people doing interesting work in the world of cloud, thoughtful commentary on the state of the technical world, and ridiculous titles for which Corey refuses to apologize. This is *Screaming in the Cloud*.

Corey: This episode is sponsored in part by LaunchDarkly. Take a look at what it takes to get your code into production. I'm going to just guess that it's awful because it's always awful. No one loves their deployment process. What if launching new features didn't require you to do a full-on code and possibly infrastructure deploy? What if you could test on a small subset of users and then roll it back immediately if results aren't what you expect? LaunchDarkly does exactly this. To learn more, visit www.launchdarkly.com and tell them Corey sent you, and watch for the wince.

Corey: If your mean time to WTF for a security alert is more than a minute, it's time to look at Lacework. Lacework will help you get your security act together for everything from compliance service configurations to container app relationships, all without the need for PhDs in AWS to write the rules. If you're building a secure business on AWS with compliance requirements, you don't really have time to choose between antivirus or firewall companies to help you secure your stack. That's why Lacework is built from the ground up for the Cloud: low effort, high visibility and detection. To learn more, visit www.lacework.com

Corey: Welcome to *Screaming in the Cloud*. I’m Corey Quinn. A recurring theme of a lot of my nonsense has been finding hapless companies who have not been adequate stewards of the data with which they have been entrusted and giving them the ignominious S3 Bucket Negligence Award. That seems to be something that isn’t well-appreciated in some areas, so I figured, let’s have a conversation about that in a bit more depth. Today’s episode is sponsored slash promoted by Open Raven and I’m joined by Mark Curphey, their co-founder and chief product officer. Mark, thanks for joining me.

Mark: Thanks for having me.

Corey: So, let’s start at the very beginning. As a co-founder and chief product officer, that means that you’re one of those folks who very early on presumably had part of the idea, if not the entire idea for what the company does. What is Open Raven, and where did you folks come from? What problem are you aimed at solving?

Mark: Sure. So actually, it’s an interesting story. I had previously done an application security company called SourceClear that I sold to CA. My co-founder Dave Cole was the early product guy at a company called CrowdStrike, which recently IPO'd. And David and I had always wanted to work together; really didn’t know what to go do.

And the honest truth is we decided to go be good capitalists and went out and asked our chief security officer friends, “What’s the biggest problem that you’ve got?” And resoundingly, it came back that, “I don’t know where my data is. I don’t know what type of data I have. I don’t know how it’s being protected. And data breaches are happening all the time, and it’s probably the big thing that I’m going to get fired for.” So frankly, Dave and I rubbed our hands together and said, “I think we can make money off of that.” And solve a meaningful problem. And hence, the Open Raven company as it is now.

Corey: Which is absolutely something that is increasingly in the public eye. Well, we’d like to hope. At some point, people just shrug, give up, assume that everything about them is public, and that’s the end of privacy to some extent, and get on with their lives. At least, that’s the negative story. I like to believe that on some level, getting better than we are today is possible.

And what infuriates me, and why I started giving out S3 Bucket Negligence Awards personally, isn’t because you wound up getting breached. I view, on some level, that is being akin to taking an outage: it happens to everyone on some level, and you have to prepare for it as best you can. All right, I get that. One of the problems that we tend to see from all corners is that companies that wind up getting breached are, in many cases, exposing data that isn't theirs, that no one consented to have handled by these folks. We see it, in some cases, with some of the credit reporting agencies and some of the data brokers. And it's not always S3 buckets, but it is the consistent drumbeat of companies not being adequate stewards of the data that has been entrusted to their care.

Mark: Yeah. I mean, look, it's certainly true that a lot of people have breached fatigue; this stuff's been going on for years, and years, and years. I think that the S3 Negligence Awards, or the Bucket Wall of Shame kind of go back down to DEFCON hacker conferences. It's called the Wall of Shame from passwords. It's not necessarily a new phenomenon.

And I would also say that whilst S3, you know, you open *The Register* and every day, there's an S3 bucket thing, it's certainly not only S3. We know that; we've been doing some profiling of things, and Elastic, and MongoDB, and everything else is hanging out there. But I guess buckets, sort of, tend to be so easy just to make them open and host data on them in the first place. But I think you're right: companies that have data, whether it's knowingly capturing it or processing it, you have a duty of care, at managing and holding someone else's data. And it just feels like people don't take that duty of care seriously enough.

Corey: And what's more, is that you'll often see a company get breached, and, "Oh, your data has been subject to a data breach." And ideally, you wind up getting that notification before you read about it in the papers. And a lot of the companies that you do business with, that contact you are very quick to point the finger of blame at a third-party contractor. Well, I didn't hire the third-party contractor. You did, and if you're not willing to wind up owning up to that, well, you're effectively trying to outsource the work—which is fair—and the blame, which is not. How do you stand on that?

Mark: Yeah. Well, it's a system that we happen to be using, but it was someone else's problem, that the default configuration was their problem. I mean also, Corey, I can tell you I have a lot of friends in the forensics industry who deal with incident response, and still to this day, the vast majority of data breaches and never reported; I know of breaches that have happened in major public companies where all the breach laws are such that they should have notified their customers and they should have notified the authorities, and it just doesn't happen. So, it's one of those problems that I think it's like the iceberg problem, right? And to a large extent, it's kind of an interesting one in that when someone notifies their customers, they're doing it from transparency.

And whilst I think you and I will both appreciate that and place more trust in those companies, the reality is a lot of the public wouldn't. And so the incentives aren't necessarily aligned up there around why, and why they should do it.

Corey: I would take it even a step further than that. I would argue that I don’t know if it’s a majority, but a significant number of breaches are almost certainly never detected in the first place. On some level the, “Oh, we’ll detect data breaches,” as a pitch that a vendor makes to a company is going to be met on some level was, “Good Lord, no. Why would we want to do that? We are happier not knowing.” And that depresses me.

Mark: Absolutely true. I’ve been building security tools for 20 years, and you’ll be surprised the amount of people that, if I deploy your tool, even as a trial and we find that we have problems, then I’m legally responsible for going and dealing with it, and I won’t touch it. The other thing that’s kind of related to that is that the security guys are incredibly busy as well, and the security tools, historically, generate lots and lots of noise; very low signal, lots and lots of noise. And so the security teams look at it and they go, “Oh, my gosh, I’m going to get a whole bunch more noise that I have to go deal with, and a bunch of more work that I have to go do. Can I bury my head in the sand?” Like, “Sure.” And it happens. That’s just the reality of the world we’re living in, unfortunately.

Corey: It is. And for better or worse, I think that it’s a world that we’re sort of stuck in to some extent. Do you think that the drumbeat of open S3 buckets that have been misconfigured containing sensitive data, it feels like we aren’t seeing as many of those as we once did, but is that just because people aren’t reporting them? Is it something that is going away slowly but surely? Or is it just as bad as it’s ever been, but it’s not making headlines anymore?

Mark: So, I’m actually building a tool to profile the AWS IP space for all the open buckets, and all of the open Elasticsearch and MongoDB thing. There’s a few of those that are out there, like Greyhat Warfare, which you can go search for S3 only, but not all the other stuff. I think when I looked up there the other day, I want to say there were about 750,000 open buckets. Now, of course, not all of those open buckets are—a number of them should be open. That’s kind of why they’re there and et cetera. So it’s—

Corey: Oh, I keep getting alerts constantly about open buckets that I have that are intentionally open, and I get alerts in the console, and I get emails at least quarterly, and this bucket that it starts with the word assets dot and then some domain, yeah. How about that? That is, in fact, designed to be open. On some level, I wind up—a few of them—just slapping a CloudFront distribution in front of it, not because I need it. Just because I want it to stop nagging me.

Mark: Of course. That’s the signal-to-noise problem again. And honestly, that’s part of the reason why Open Raven’s doing very well is that all these companies have been hiring people to go around, chase down open buckets, which are designed by functionality to be open in their organizations but don’t contain any sensitive data. So, I don’t know. Look, to answer your question around the S3 bucket thing, I think again, like breaches, people have got a bit of fatigue.

And there’s only so many articles *The Register* can do with, “Hey, S3 bucket open to the world,” and—what is it—27 terabits of data or 900 terabits of data. It’s not necessarily one-upmanship on the headlines anymore. But the gut tells me the faster we deploy things, right, the whole kind of DevOps movement, in many ways is moving against the security grain. And rightly so. So, to think that the problem is getting better, I don’t think is reasonable.

And then I think that Amazon, in some ways, have done good things with the security policy and all of those types of things, but they’re largely designed for greenfield environments. And when you step off the reservation, I mean, gosh, what, do you think the average developer is going to be out of figuring out how to configure that XML policy in an average XML bucket? Of course not. They’re just going to make the damn thing open, and they’re just going to move on and do their job. So, we’ve got to figure out how to get better tooling, easier, secure by default. And then, like you said, we’ve got to figure out ways to reduce the noise so that people can act on signal, and not get bombarded by noise, and just shut it down.

Corey: My approach to cutting through that noise on open S3 buckets, as I tweeted out a couple of times now, is to just copy a few petabytes of data into the open buckets. My operating theory is that while you’re going to ignore a politely worded email from a security researcher, you’re probably not going to ignore a bill that is 80 times larger than it’s supposed to be at the end of the month. That seems like it might be—among other things—legally fraught. What’s your approach at Open Raven to solving this particular problem?

Mark: Well, so for us, it’s all about improving the signal-to-noise. So, it’s about setting up a policy that you can say, “On this bucket, this is the type of data that I’m expecting;

these are the security controls I’m expecting; if it deviates from that in any way, go, let me know.” And then we use OPA, Open Policy Agent, to go check that and go send alerts or pump stuff out to a firehose, pump it to a security event system, or whatever. So, in general, that’s it. It’s like, define what goods meant to be, and then let me know when good is not occurring so I can go figure out how to deal with it.

And of course, you know, you create generic policies like, “Look, I never want to see financial data on a bucket that’s open to the Internet and that’s unencrypted,” or, “I never want to see something with a CIDR range of 0000 and through some security group,” or something. So, that way, essentially, what companies get to do is, sort of, encode their intended use policy and alert when that’s not there. So, for us, it’s all about that. Part of what we see, and I’ve seen this in the security industry for the last 15 or 20 years, it’s there’s textbook security and then there’s the real-world security. You go look at a lot of, kind of, textbook security solutions, and they’re fine.

They work absolutely fine. I worked at Microsoft for a long time, and I was always amazed at how everything worked perfectly at Microsoft, and then when you stepped off the reservation, nothing worked properly. But everyone in Microsoft would be scratching their heads going, “Well, it all works fine here.” It’s like a developer saying, “Hey, it works on my laptop. It was only when I committed it to CI the problem occurred.” So, it’s that same thing, and you got to design stuff for the real world.

Corey: You also say it goes beyond just S3 buckets, which I believe. For a while there, I think—was it Elasticsearch, or was it Mongo that had a default password of ‘changeme’ or something horrifying like that?

Mark: Yeah. One of those. I forget which one? It was? Elastic’s a big offender, for sure. Mongo is a big offender, for sure. But I mean, you also see, like, Jenkins servers that are sat out there. I mean, that camera thing—what was it—the Verkada thing recently? That was a CI server that happened to have a script that had access to loads of things. But the amount of Jenkins servers that are accessible through the internet is shocking. It’s not just buckets for sure.

Corey: It definitely becomes a weird thing. I don’t know if there’s a fix here—I really don’t—longer term. But instead of looking forward for a minute, let’s go back and visit the past for a bit. You were the founder of the OWASP reporting list. What is OWASP? Is it a list? I’m most familiar with the OWASP Ten. But I’m certain you’ll have a better story on that than I will.

Mark: Yeah, no, no, no. Top Ten was this whole thing. So, I was running software security at Charles Schwab, early 2000s, 2001. Before, it was kind of a really big thing. And we used to get vendors coming in trying to sell me products.

And honestly, it was kind of a joke. My market open would have 8 million accounts, like, a trillion dollars under asset, and people would come in and try and sell me a web application firewall, which, maximum throughput was like 0.01% of my market open traffic, and things. But there was nothing out there on the internet to go point to and to say, “Well, this is good.” It was basically me versus a vendor coming in.

And so I said, “Right. This is kind of crap, right?” And I got together with a bunch of other people that were also doing similar things, some other people at some other banks, some other people in other companies. And I said, “Right. I’m going to go publish something.” And I wrote it over a weekend, literally wrote this guide, called the “OWASP Guide.” And it was basically a set of principles around software security, like lease privilege—you know, nothing sophisticated, but it was those types of things—and published it.

And then OWASP was basically born. So, it’s the Open Web Application Security Project. Then it got a lot of traction because a lot of people had signed up, and a lot of people were then starting referencing this to build their own application security programs. Over time, of course, OWASP got very successful. I think it’s, like, 40,000 people or something like that turn up at those conferences all around the world and chapters all around the world.

And there's lots and lots of projects that have taken place, one of which is the Top Ten that you referenced, that a lot of people know of. And the Top Ten has been around I want to say since, like, 2004, or something like that. I don't know, I'd have to go back and check with history. It's hardly changed since 2004. And you can have a good conversation around why that is. But yes, that's the history of OWASP.

Corey: And now, of course, you have this list that doesn't seem to have changed significantly in a while. I mean, back when I was starting up the *Meanwhile in Security* podcast and newsletter with Jesse Trucks, we talked about that being one of the key problems is everyone wants to know how to handle security in cloud, but if we take a look at how a lot of application vulnerabilities exist, that list hasn't materially changed. If anything, the advent of cloud has fixed some security issues, in that you're not allowed to muck with them anymore. Datacenter physical security is no longer a vector for most folks who are all-in on a public cloud provider. But you're also dealing with this other problem of, where, now it's a list of enumerated S3 buckets, for example, and if you misconfigure that, it's something that's globally known, and I guess it removes the security-through-obscurity argument, insofar as ever was one. Has things changed in a time of cloud or is it just the same thing with new labels on it?

Mark: Well, I mean, there's a couple of things. So, you've got to ask yourself, what is the OWASP Top Ten of, right? Is it the top ten most popular issues? The top ten most severe issues? The top ten voted by security people?

Like, no one's ever really been able to get to that, apart from an arbitrary top ten. And I don’t want to take anything away from it because the Top Ten has been incredibly useful in getting to developers, giving them a tangible, like, ten things; go focus on these ten things and you’ll raise the bar. So, that’s kind of piece number one, but has it changed? Well, should you have expected it to change depends on what you believe it’s based on. If you go look at them, though, like, no.

Things like injection, and broken authentication, and sensitive data exposure, those things haven’t changed because they’re just general things and they’re going to be around forever. You think sensitive data exposure is going to go? Doesn’t matter what technology we change, it’s always going to be there. For me, though, what’s kind of interesting about it, and why maybe I’m a bit of a skeptic about it is that you can eradicate total classes of problems—I believe—by changing patterns. So, a good example is, look, if you go use one of these modern development frameworks, application frameworks.

It’s built-in inherently. And the same with a lot of SQL injection problems that you used to see all over the place. You’d have to intentionally go create those problems, for the most part, now. And I think the cloud is done the same. It’s taken a lot of problems away, it’s extrapolated them into a service, it’s extrapolated them into a pattern, and the pattern can then go away.

So, back to the S3 thing, I think there's hope [laugh] because if you can make a change upstream, I mean, you've probably seen recently all these damn, you know, supply chain attacks. And the bad guys are going further and further upstream where they can affect things downstream. And the good news about all of that is if you can figure out upstream, the way to go secure it, everything downstream gets secured as well. So, I think with a lot of these things, if we can, instead of trying to play whack-a-mole or put the finger in the dike, if we can start thinking about patterns and ways to go solve them at a class or problem level, then we stand a chance of fixing them.

Corey: This episode is sponsored in part by ChaosSearch. As basically everyone knows, trying to do log analytics at scale with an ELK stack is expensive, unstable, time-sucking, demeaning, and just basically all-around horrible. So why are you still doing it—or even thinking about it—when there's ChaosSearch? ChaosSearch is a fully managed scalable log analysis service that lets you add new workloads in minutes, and easily retain weeks, months, or years of data. With ChaosSearch you store, connect, and analyze and you're done. The data lives and stays within your S3 buckets, which means no managing servers, no data movement, and you can save up to 80 percent versus running an ELK stack the old-fashioned way. It's why companies like Equifax, HubSpot, Klarna, Alert Logic, and many more have all turned to ChaosSearch. So if you're tired of your ELK stacks falling over before it suffers, or of having your log analytics data retention squeezed by the cost, then try ChaosSearch today and tell them I sent you. To learn more, visit chaossearch.io.

Corey: I sure hope you're right. I mean, in an ideal world, you will be. But it's, ugh, I have so much trepidation [laugh] around all this. And I don't know how it's going to wind up playing out. And I hope that it's going to go well. But it just feels like you're constantly railing against the tide. And I don't know how to wind up addressing that. I really don't. I wish I did.

Mark: Mm-hm.

Corey: Is there anything you can say that helps them be more optimistic about this, at least?

Mark: [laugh]. Well, I mean, you're right. Look, I'm no longer in the application security business after spending 15 or 20 years in there because I just gave up trying to convince developers to care about security. I just—and I don't blame the developers. They've got another job to go do and security's too hard.

So, for me, it was just pushing molasses uphill. And, I think, to your point, yeah, why would you expect anything different if we carry on doing the same thing? And the reality is, we're moving faster and faster, we're making it easier and easier to deploy things, we're getting more and more complex systems. Why would you expect anything different? So, yeah, I don't think you're skeptical for a bad reason.

Corey: No. For better or worse, we still wind up having these problems. I don’t know how to solve it. I really don’t.

Mark: I mean, look, for the reality, if you go back to the old days, like, the old school—obviously I’m a bit of an old person, right—but you go back to some of the military things used to be, like, “Trust, but verify.” That motto works incredibly well. You trust people are going to do the right thing, you verify they’ve done the right thing. That means you don’t hinder the speed, but you go back and check and if anything happens, you come back. And it’s like, accepting things.

One of the other ones around that was, like, it’s people, process, and tools. People, process, and technology. And again, technology is never going to solve the problem of security. It’s a people problem. “You can’t patch stupidity,” and all of those phrases.

But if someone gives someone access to a local root account, or whatever the thing is, doesn’t matter how many other security controls you’ve got. I mean, I’ve seen it in cloud environments, as I’m sure you have. Someone goes and creates a security group, 0000 so they hop in the thing from home and don’t have to come in and go through all of the other control points. And it’s just the way stuff works. So, if you have that—if you take that mentality of, “People, process, and technology,” and, “Trust, but verify,” I think, use the right technologies and build the right process around it, then you can at least manage the risk. The risk is never going to be zero, but you can at least manage the risk to an acceptable level.

Corey: Let’s pivot a little bit and talk about the flip side of data security. And that comes down to privacy. There’s been a bunch of regulatory efforts around that. GDPR, for example, California has its own version of that that’s going out, and there’s also a growing school of thought that thinks, on some level, we’re post-privacy. Where do you stand with that?

Mark: Yeah. I mean, look, the privacy regulations are raging right now. You got GDPR; you’ve got CPRA, the California one; you got HIPAA, the Health Information Privacy Protection Act. And they’re all over the world. Japan has them, Australia has them.

They’re all over the place. And I think the US now is talking about having a central breach law around privacy data. The great challenge is that we’re all becoming a data economy, and companies are all becoming data companies, and so they want to gather more and more data. And the reality, I think, is that this whole stuff around cookie consent, I just think it’s just nonsense. When was the last time you said, hey, I’m not going to consent to you using my cookies?

It’s kind of like back in the old days, when you said, “Hey, I’m not going to allow JavaScript to run in my browser.” Like, all of a sudden, nothing works. And you’re like, “Oh. I’ll succumb.” But then before you know it, data it’s been over-reached, right?

You probably saw the Alexa the other day that has the radar so it can watch you sleep in your bed. Sure, of course, they're not going to use that data for anything bad. But next time a breach happens, or some clever data science person decides to correlate something—I don't know what it might be—in the middle of the night, it happens. So, I think what you're starting to see is that you're starting to see regulators and legal people who don't really understand technology, regulating to prevent those bad things happening. And then technology trying to figure out how to go and meet those regulations, but meeting it with the absolute minimum bar versus trying to figure out what the actual intention is.

And I think you're going to see a bigger and bigger gap. I mean, look at what happened with third-party cookies as an example. The whole third-party cookie thing we saw, what was that the CORS headers, we saw anti-cross-site scripting headers because all of those things started happening. And then what does everyone do? They just go call a tracking pixel.

And then all the marketing automation tools carry on working as possible. So, I mean, I think you've got a balance between technology working as intended in certain good use cases, and there are people using that for their own use cases, which either break or push over the line of privacy. I don't know. How do you see it?

Corey: I think on some level, it’s not necessarily that people care necessarily that some company in the aggregate knows what they’re doing. There are some that do, and I’m not disputing that. But for most of us, I don’t necessarily care if Google, for example, knows what I browse on the internet. I care much more if you—personally—know what I—personally—am browsing on the internet. So, there’s a question of, once they have that data, do I really care that much about what they do with an aggregate? Not really? Do I care what they do about it on individualized basis? Kind of, yeah. And do I care if they’re making, then, that individualized data available to third parties? Absolutely.

Mark: Yeah.

Corey: It comes down to what the use of that thing is. Now, I know that I am not going to win friends with that particular argument myself. And I get it. In an ideal world, I think that advertising should be something radically different than it is. There are advertisements in this podcast, for example, and they’re catering to an audience that cares about the topics we talk about on this podcast.But I have no tracking data of who listens to this, other than raw download numbers and rough GeoIP by continent. It’s not something that is ever going to be attributed—at least from where I sit—to individual listeners, nor would I want it to be.

Mark: Yeah. But, look, here’s where I might be able to convince you otherwise of that. In China, there is a well-known place called the Beijing Genomics Institute, and the Beijing Genomics Institute do genetic engineering, and not necessarily for good. So, it’s not necessarily to find cures for things, it’s also for other nefarious purposes. And the

Beijing Genomics Institute acquire DNA data from US hospitals, US healthcare systems when you get your blood checked.

Now, that data is supposedly aggregated, but once you can start pulling apart DNA strands. You can start identifying people at different levels. And I think that's the danger. There's been a lot of cases where de-anonymizing information is possible. And so you're making the assumption that that data is generally de-anonymized and use for the right reasons, but there's been case after case where that's not the case. So, maybe you'll change your mind on that, Corey. I don't know.

Corey: Maybe. I also, on some level, feel like I'm fighting a losing battle against the tide.

Mark: Yeah, yeah. My wife says, "Aren't you worried about your credit card going missing?" And I'm like, "I'm sure it's in many, many databases at this point." I rely on Visa, at that point.

Corey: Well, that's also a separate problem, too. I mean, this idea of, "Oh, your identity was stolen because someone else has opened a credit card in your name or stolen your credit card." My very honest response to that is, "Oh. So, you weren't cautious about who you decided to lend money to and validate they were the person you thought. And you're trying to make this my problem because why, exactly?"

Mark: Yeah. I mean, look, in those cases, and that's why it's the corporate's responsibility to deal with those issues. I guess it's the same with social security numbers, in that they're out there in so many places on the internet, and they're pushed around in so many different ways, aren't they? I think we've got to start moving into some of these zero-trust kind of protocols, and zero-knowledge ways, and all of that type of thing and the future.

Corey: Indeed. And I think that there's one thing that every corporate entity listening to this—or representative of same—can agree on, and that is they prefer this conversation to remain hypothetical and aimed at the abstract not at them right after they've had a data breach, which of course brings us back to Open Raven and how it aims at these things. You do have a—at the time of this recording, it is still upcoming—a paper coming out contrasting what you have built with I believe it's Amazon Macie?

Mark: Mm-hm. That's right. Yep. Yep. So, when Dave and I founded the company, we went out, like I said, and we asked everyone, what's the biggest problem, and it was data security. And then when you broke that down, it broke down into, "Let me know where my data stores are." So, do I have buckets? Do I have stuff in RDS? Do I have stuff on file systems, et cetera? "What type of data do I have there?" "How is that data being protected?" You know, access control, and encryption, and all that things, and who has access to it?

So, it basically broke down to those things. Those things haven’t changed at all. So, think of that piece number two—what type of data do I have—as being data classification. Amazon have a service called Macie, which works on S3. So, we’ve built that feature.

Now, lucky for us, as it turned out—you get few really good breaks in the startup world—is that Amazon Macie it turns out it’s not very good, and incredibly expensive, and very, very slow. So frankly, the way we market it is, “Cheaper, faster and better than Macie.” And we believe in transparency of that. Every vendor will say we’re way better than everything, right? So, we’ve kind of done what you would do with a clinical trial in that we have basically built a—you know, here’s the test.

Here’s exactly what we’re going to test for, kind of like, laying it out in an academic paper. Here is the data, so you can go rerun the test yourself. And here are the results. And we know that we are way, way, way more accurate than Macie. We’re deployed as Lambda functions so we can scale up and run much, much faster than Macie. And then, certainly way, way cheaper than Macie, but that wouldn’t surprise you at all in that case would it?

Corey: No. Even after their massive recent price reduction, it was still, okay. That is in fact, still incredibly expensive, across the board. I mean, my argument with the original Macie and its pricing was I had a customer at that point, eyeing it and doing some math and, yeah, okay, first month would have been $76 million to run it in their existing stuff, which was significantly more than at that point, their annual AWS bill. So, it was, “Okay, let’s go with Option B,” which is literally anything except that and you’ll save money.

Even a data breach wouldn’t have been that disastrous compared to the pricing story. And now they’ve cut it to 20% of that, but that’s still an eight-figure bill to run these analytics on their data set. And that is… that’s not tenable. And on some level, it becomes the differentiated value of doing that isn’t there for customers. If I wound up running all of the various security services that AWS offers on an environment, it’s pretty clear that it would cost more than the data breach would.

Mark: Well, it doesn’t even work. Even if the cost thing was put aside, one of our customers tried it. I think they spent a million and a half on a trial, in a month, and it found 30 first names in a credit card database. I mean, it’s kind of crazy. And when you pick it apart underneath the hood, it’s a giant regex, essentially, and just doesn’t really work.

I mean, the reality is that that thing was built—it was actually a—it was originally an In-Q-Tel project, which is the funding arm of the US intelligence agencies. It was called [Harvest IO 00:28:04]. It was an acquisition that they bought in. And it was built a long, long time ago. If you want to do data classification today, you have to be able to not only identify structured, unstructured, and semi-structured data, and it comes in all places, and it goes into all file formats, in S3 buckets, it’s Parquet files—which are the backend of LakeFormation and Lakehouses and things like that.

But when you find a piece of data, you’ve got to be able to go and validate, is that data real? I mean, take an AWS API key as an example. It’s very easy to go figure out how to push that thing into that format, but is it a real key? Whereas if you use validators, go login to an AWS API and you’ll get a return that will say, “Is this a valid key, and which account is it associated with?” And so we’ve done, both in terms of the accuracy of identifying the information stores, the tests that we’ve got show, in general, we are twice or three times more accurate than Macie on finding the initial piece of data.

But then we have these validators. So, you know, you get a credit card, go call a credit card API. Is it a real credit card or is it just a 16 digit int? And you can go check that stuff. Data classification has moved on since that stuff was there.

So, even if the pricing thing was fixed—and as you point out, it certainly isn’t—it’s just not a good option for people. And then the kind of second piece to that is that the majority of customers that we see, and people, are looking at things like Snowflake. I mean, if you look at these data platforms, Databricks, Cloudera, Snowflake in particular, you know, they’re built on top of AWS services. But people are moving data to those places, so it’s not just an S3 problem, as I said. It’s about people putting data in Elasticsearch, in RDS, in file systems.The data is everywhere—and backups. Like, all of this stuff gets pushed up into backups and stuff as well. And so you’ve got to have a service which goes and checks it. We decided to go compete with S3 and beat Macie first, but that’s certainly not where the tool and technology is going.

Corey: No, for better or worse, it would seem not. Thank you so much for taking the time to speak with me. If people want to learn more about Open Raven, what you’re doing and how you’re doing it, where can they find you?

Mark: Yeah. openraven.com is the best place to go. We’ve also got a pretty exciting open-source tool coming up soon, which is called Magpie. Magpie is a cloud security posture manager.

So, think of it as we’ll go out and check all of the security settings on your AWS environment. And so we’re releasing that open-source around the end of April as well. So, keep an eye out for Magpie. We’re taking the core out of Open Raven that does all the discovery across the orgs, pulls back all the attributes, or the IAM, or the security groups, and then allows you to go write security rules on top of that. Not data rules, which is what the Open Raven platform does, but security rules. So, also go check that out, but all linked off of openraven.com.

Corey: And we’ll of course put links to that in the [show notes 00:30:43].

Mark: Wonderful.

Corey: Thank you so much for taking the time to speak with me today. I really appreciate it.

Mark: No, thank you very much, Corey. Much appreciated.

Corey: Mark Curphey, co-founder and chief product officer at Open Raven. I’m Cloud Economist Corey Quinn, and this is *Screaming in the Cloud*. If you’ve enjoyed this podcast, please leave a five-star review on your podcast platform of choice, whereas if you hated this podcast, please leave a five-star review on your podcast platform of choice along with a comment enumerating all of the S3 buckets you have inadvertently left open.

Corey: If your AWS bill keeps rising and your blood pressure is doing the same, then you need The Duckbill Group. We help companies fix their AWS bill by making it smaller and less horrifying. The Duckbill Group works for you, not AWS. We tailor recommendations to your business and we get to the point. Visit duckbillgroup.com to get started.

Announcer: This has been a HumblePod production. Stay humble.