MALWAREBYTES FIGHTS THE WAR AGAINST MALWARE WITH BIG DATA

In this Run and Reinvent podcast, technology evangelist, Joe Goldberg speaks with *Darren Chinen, Senior Director of Engineering at Malwarebytes, who talks about winning the war against malware with big data. He has vast experience in Data Engineering and has led teams at companies like Apple and GoPro in the past, helping them turn data into valuable insights.*

Jill Perez: Welcome to the BMC Run and Reinvent Podcast. I'm your host, Jill Perez, and today's guest is Darren Chinen, Senior Director of Engineering at Malwarebytes. He has vast experience in data engineering and has led teams and companies like Apple and GoPro in the past, helping them turn data into valuable insights. We're also joined by Joe Goldberg, technology evangelist. He is here at BMC. Today we're gonna be talking about winning the war against malware with big data. So, Joe, I'll let you take it from here.

Joe Goldberg: Hey, thanks very much, Jill, and welcome, Darren. It's really great to have the opportunity to speak with you. I think this is a great topic that everybody is gonna be interested in. You know, Malwarebytes is a company that I personally have been using, and I think lots of people know all about it, but maybe you can tell us a little bit about Malwarebytes as a company and the mission that the company was really founded around.

Darren Chinen: Well, thanks for having me here, Joe. Yeah, this is a really important piece of our history, which is how we were founded. Our CEO and founder, Marcin, was 14 years old, playing video games, and caught a virus on his computer, didn't want to tell his parents about it, and he went

searching on the Internet for some people that could help him clean his machine. And so, at the ripe old age of 14, with the help of some people he now calls heroes who are actually still working at this company, he was able to write his first Malwarebytes program, and that kind of evolved. And eventually, by the time he was 18 years old, he was able to found Malwarebytes. And now he's — I think he's now almost the ripe old age of 30. I can't believe it. He's getting old.

And really, our mission has stayed the same from the very beginning. We absolutely can't stand malware. We imagine a world without malware. And what I think is the best thing about what we do is we always talk about if there's no malware in the world, there's no reason to be in existence anymore. So, I don't know if we'll ever be able to achieve that goal, but it would be an amazing goal to eradicate the world of malware and no longer be a company one day.

Joe: Well, that is certainly interesting, some background about the company that I wasn't aware of. But I don't think you have any fear whatsoever of putting yourselves out of business. I think there seems to be an ever-growing army of folks out there who are very much trying to encourage and drum up business for you. So, I understand you are leading the data engineering team there. Could you maybe give us a little but of a description about the role of that team within Malwarebytes?

Darren: Sure. I lead, actually, three separate teams here. The first team is the team that actually does the website, what we call the "dub-dub-dub team." I have an infrastructure team. And the team that we'll kind of double-click on today is the data science and engineering team. And that team really handles all of the big data. We have some anonymous telemetry streams coming in that pours out all kinds of IOG data. Think of it as sensors all around the world. And they also do some of the data science and AI as well.

Joe: Wow. So, one of things I think, certainly most people don't draw connection — certainly not immediately — between malware and malware detection and big data. And some of the things you're describing that your team is focusing on are among the most interesting, I think, topics that really are a focus for almost every organization nowadays. Just pick up the newspaper, and everything you hear, everybody is trying to extract value out of data, and big data is really top of mind for everybody. Can you maybe talk a little bit about how Malwarebytes is leveraging the power of the data that you're collecting?

Darren: Sure. It's really been kind of serendipitous, I think. Marcin had this — Marcin, our CEO and founder — had this vision of everyone has a right to a malware-free existence. He's been giving out our product for free — even to this day, you can download it for free. And one of the things that we do is we collect anonymous data that act as sort of sensors out there for us to understand the landscape of what's happening in the world and where are attacks happening.

It's really interesting because, Joe, if you were to look at some of the telemetry years ago, the world of malware was evolving at maybe a monthly, or sort of a slower than monthly pace. And today, sort of with modern technologies that help software engineering, things like CICD and continuous integration and delivery, the world of malware is literally changing by the hour.

So, it's very, very, very important that we're able to collect and harvest and mine all of this data in an efficient manner so that we can understand the landscape, and our company and our systems can adjust appropriately to combat malware as it's involving on literally and intra-day basis.

Joe: So, as you're talking about this, I'm thinking to myself, you know, you mentioned CICD, that certainly is another technology practice that's on top of mind for many organizations as a way to deliver new business services and to serve their customers. It's really interesting to think about how

malware creators are using agile software methodologies to deliver malware more quickly.

Darren: The world-class software engineering organizations, unfortunately — the television shows you see with guys in dark rooms and sunglasses and a big bag of Doritos next to their desk is — that's probably far from the truth. These are very, very sophisticated software engineering shops that are in business because there is a path to cash, and they're making money off of things like ransomware. So, absolutely, for them, it makes sense to stay on leveraging the most cutting-edge technologies. And really, to fight the war against malware, we have to be as good as or better than whatever they're doing.

Joe: So, that's a really interesting concept, that you're sort of engaged in this battle of — and not just you, but the entire industry, I think — in this battle, using the latest techniques and technology, on the one hand for, perhaps, evil and the creation of malware, and on the other hand, using that same set of technologies and processes and new ways of think to try and combat that.

So, recently I heard you talk about how your team is not just using AI and machine learning, which again, I think so far our conversation has been chock full of things that are really top of mind for everybody. But, not only are you leveraging these cutting-edge technologies and approaches to building, I guess, malware detection, but to be able to do it at scale and to industrialize that. Can you elaborate on that and maybe share with our listeners how you're doing that?

Darren: Sure. Well, first of all, make no mistake about it, right? I mean, the bad guys are using AI just like we are. So, it's not like this technology, this open-source technology, is exclusive for the good guys. The bad guys get access to all of it as well. So, for us, it's all about — we have various layers. I think there are seven layers of protection in our primary endpoint protection product. Not all of them are AI; some of them are other techniques that we use. AI is one type of technique, a very effective technique that we use. And we use it both in our endpoint technology as well as — our behavioral EPR technology is coming out with an AI engine as well, and we've been testing that.

So, one of the things that we recognize is that we can build sort of anomaly detectors with AI. AI is actually decently good as anomaly detectors, as long as you can narrow the problem set. And one of the problems with malware is that you just have this — well, I should say, not with malware, excuse me — one of the problems with machine learning is that it's a machine, right? So, if you think about machines in general, they're very good at specific tasks and specific problems.

So, when you walk into your kitchen, you have a machine that is called a toaster, and all it does is toast, toast, right? And you have another machine called a waffle maker that only does the waffles, and you have another machine that's called a blender, and it just does blending. Why isn't there one machine — or the Jetsons' robot, right? — that actually does everything? The answer is because machines are good at very specific tasks. It's actually really, really, really difficult to build machines that sort of generically do everything in the sense that a human can.

And so, it's funny, because I was talking with my brother at Google long ago, and he was giving me a little a talk, because he does machine learning and he's in the research group over there. And he says — you know, we joke, because we say the difference between identifying — building an AI model to identify the Eiffel Tower in a picture and an AI model to identify everything in a picture is sort of the difference between maybe two weeks of work and a lifetime achievement. And so, that's one of the problem statements, and let me bring it back to malware.

When you're trying to figure out if a machine is infected, you have to look at the behaviors of that machine, or that's one of the techniques that you can use. Now, the problem is that the way people

use machines or PCs today is wildly different. You can imagine a salesperson uses their machine in a much different way than a marketing person, and the person who really messes up their machine is probably an engineer or a dev-ops guy, right?

And so, if I try and figure out what is anomalous behavior, and I'm just looking at the machine logs of the activity of what's happening on these machines — if I'm not doing this correctly or if I'm not careful, I'm going to either be too sensitive, and everything that the engineer and the dev-ops guy does, I'm gonna think, oh my gosh, these guys are infected with malware, or I'm not gonna be sensitive enough, and I'm gonna think that the engineer and the dev-ops guy is normal behavior. And when the sales guy is experiencing the hack, I won't recognize it, because it will look just like an engineer.

So, it's a signal to noise problem, and that is sort of one of the key things that we have to cover in the world of malware detection. We have to improve the tradeoffs between wanting to do a good job at detecting malware, but also not making a mistake. We call that a false positive. And if you're in the security industry, a false positive is what we jokingly call an extinction event. It's one of those things that can put you out of business. So, we have to be very careful about getting good detections, but also not making catastrophic mistakes.

Joe: I certainly have personally been, I guess, a victim — it certainly wasn't an extinction event, but I can appreciate false positives with my credit card. And, in fact, on a few occasions after my credit card was no longer working, I called in and was told that I was behaving just like a typical fraudster when I was going about my daily life. So, certainly, if you expand that to a broader perspective, that could certainly be problematic.

So, I think that the problem that you're describing, the challenge that you're describing — I'm assuming that one of the big things that can help is how you leverage technology and what your underlying architecture is that supports the kinds of processing that you are describing. So, maybe you can talk a little bit about what does that look like at Malwarebytes. What kind of technology are you using, what's your architecture like? Some of the things that help you avoid false positives and to identify it from that signal that you're seeing and separating it from the noise of normal activity?

Darren: Right. Well, I think there's a couple of phases to this. The first one is how you harvest the data. So, there's a lot of different types of data that are coming in. So, we have this IOT type of data, what we call telemetry data, and that data really just gets a public-facing API. We quickly push that data into stream processing — we use Kafka and Kafka Streams. And then we're able to process that data, and then we leverage everything in a public cloud, so we can leverage existing cloud services like AWS and some of the things that they have, like S3 and ephemeral processing to process that large amount of data.

There are other types of data that come in as well, and data that we need to harvest from external APIs. And for that, we use sort of a Java framework that can go out there and harvest data from some of the APIs that we need to get to, to enrich a lot of that data with some of the other — I would call it more transactional type of data. We do some caching that happens in Redis. And then, basically, our goal is to provide a platform for the data scientists to actually go do their work. And that is one of the biggest problems, I think, in data science and AI, is that you run into a couple of issues, right?

A data scientist walks into a door, they go, "All right, it's my first day, I've got my environment set up. Where do I go to get the data?" And the first roadblock that they come up against is, they go, "Well, the data is sitting all over the place." So, the first thing you need to do is to centralize the data, and

we've done that. Then they start going and digging around the data and it became a legal issue. And it's a legal issue because of things like PII and GDPR rules and just being sensitive to making sure you treat your customers' data with respect and privacy.

And so, we go through a process of cleansing that data so that it's ready and there's a playground available, almost like a sandbox, available for the data scientists to really sink their teeth into. And then it's about tooling, right? They need a certain amount of tooling to go and do whatever kind of machine learning they're gonna do, whether that's R&R server or spark clusters and all of the notebooks. And once they build that model, they need a way to go back and sort of deploy this model from a pipeline from, I have my model and I need to give it to the data — excuse me, to the engineering organization. I need to deploy it into production and I need to get predictions out of it, right? In our case, we have models that get predictions every second, and we also have models that get where we need daily predictions. We can do it in batch.

So, there's kind of a pipeline that needs to occur between the data scientists and the actual production development engineering. And keeping those models trained in a production in environment is actually a critical task to ensure that the environments are flowing and working properly and continuing to make accurate predictions.

Joe: So, that's an interesting point that I'd like to maybe dig into a little bit more. Certainly, a lot of people when they think about data science and big data, there's a lot of emphasis on coming up with clever algorithms and going to extract learning and the — all of that kind of real-time processing and maybe doing everything just as data is flowing. But it sounds like you are describing a lot of different processes, like training models and things that have to be done in a repetitive kind of fashion. That sounds like this other sort of periodic, maybe long-term orchestration kind of tasks that have to be done. Is that right? Can you maybe talk a little bit about that component of the engineering and the delivery process?

Darren: Absolutely. It's a piece of the — it's a piece of AI that's, I think, overlooked by a lot of people. What happens is, scientists come up with this model and it works. And then when they throw it over the fence, it works perfectly fine. But over time these models, the underlying data changes and the prediction accuracy goes down. And so, you'll find data scientists sort of looking over their shoulder, and then they have to say, "Oh my gosh, my accuracy's not as good." And then what they do is they have to go retrain their model manually and give another model and it's a big deploy.

We found that process is something that we can automate. And it's about keeping our production AI environment up and running smoothly and constantly making good predictions based on the ever-changing data. The analogy, right, is if I built an AI model around celebrity detecting, and I fed it all the celebrities — let's say we're in a time machine and we're back in 1990. And I fed all the celebrities from 1990 into the model and it trained on it, it would make good predictions in 1990.

But fast forward to today in 2019, if I try to run that exact same model without feeding any new data, it couldn't accurately predict anything, right? It would say things, like, you'd feed in Pink and it would say, "No, that's a color. That's not right." Not really realizing that that's actually an artist, right? So, it's very important that you take as good or better care of your production AI deployments and the ever-changing data under there as you do with taking care of sort of building your data science labs.

Joe: Well, that's good guidance and something we can take away. Can you talk a little about maybe specifically what tools and what components help you achieve that level of automation to make sure that your models are operating properly?

Darren: Sure. Well, we use Control-M, and Control-M has helped us to orchestrate everything beautifully. We've been using Control-M, too, for all the big data. We also use Control-M on Snowflake for some of the structured data. And it really, beautifully coordinates all of our ETL processing, the batch processing, all of the ingest, sort of what we call the pre-feature bills. And then what it does is — we actually take all of our AM models, and what we try to do is get them scheduled, and if it can be done in a container, we'll definitely do it in a container.

Otherwise, what it will do is it will kick off a job on Spark, and once that model is retrained, it's tested, it goes through an approval process where we take a look at the confusion matrix to make sure that the accuracy has, in fact, improved with the new training, and then there's a process to actually promote that improved model into production. So, that's something that's pretty much handled — the backbone of that is pretty much handled by Control-M and that whole orchestration and the scheduling of all of that. That's all done in an automated fashion.

Joe: Oh, well that sounds — that sounds pretty cool. So, in that description, you mentioned yet another piece of technology, if you will, that's proving to be extremely popular, which is containers and containerization. Do you see that as something that is going to part of your architecture and that will fit into strategy, specifically with orchestration and some of the capabilities that you've just been describing going forward?

Darren: Oh, it's absolutely critical. Because once models are created and we need to deploy them, a lot of times we don't know — there used to be a tremendous amount of coordination between the data scientists and what packages they'd use and what languages, if they use Python, what version they're using. And they all like something different, right? And so, it was very confusing. And so, what we've gone to is more of a containerization model.

So, we just say, "Look, give us the container." And we'll go ahead and — almost like an API — what is the contract of what the container expects. And then we'll spin up the container of what is the contract of what the container expects and what is the output. So, the container model, for us, is really the model that's really allowed us to provide more of a data science and AI platform, as opposed to being so tightly integrated and having to have these long whiteboard sessions of exactly what components and how to keep this updated and how do I deploy this exactly and how do I test this. Having the models containerized has really been critical for our success.

Joe: Oh, okay. Well, that sounds amazing, and one of the things from our talk so far that I take is pretty much every new kind of technology that I'm aware of, that we're hearing a lot of buzz in the industry, you and your team at Malwarebytes is leveraging that and seems that it's one of the — maybe the arrows in your quiver that maybe is helping you stay abreast and hopefully stay ahead of those bad guys. Can you maybe — can you talk a little bit about what you see for, let's say, the next few years, in terms of really cutting-edge use cases and some new technologies that you're gonna be working on, I guess in the next few years?

Darren: Well, we have a running joke in our data science and engineering organization. You talk about how you try to remain agile and test new technologies. The joke is that the architecture is only good for the next two weeks until the sprint ends. So, we definitely are changing things up relatively quickly. I think over the next three years, you're gonna see AI continue to evolve. If you roll back the clock years ago, AI was something that was — you either had to be in the insurance industry, doing sort of data science on insurance, or, separately, you had to be more of a computer scientist, doing things like map producing type of stuff.

And what has happened is that we sort of evolved into sort of two skill sets: the machine learning experts that are creating algorithms; and the data scientists, who are implementing these algorithms that are more like statisticians. And I think you're gonna continue to see AI evolve the pipelining models that are being built, just that getting democratized. There's tons of analysts who want to jump from being a data analyst into being a data scientist, and I think that's gonna pull the industry in a way that makes AI more and more accessible to the average data analyst each and every year.

The other thing that's happening is that we're really moving very, very quickly towards streaming and real-time data and event-based automation. So, I think that'll be a big thing. As we move toward event-based automation, I think that's going to be big in the future. We certainly do a lot of stream processing, and of course when we see a big attack, we can't wait for a batch. We actually are gonna be doing things like event-based automation, maybe even to retrain our models, if the data is changing in real time. So, I think that will be another big focus area.

And I think you'll just continue to see the industry evolve, right? We're certainly not at a plateau. I don't know if you'd agree with that, but at least from my vantage point, I feel like we're not at a plateau yet, and there's quite a bit of room for growth, even within a mature industry like security and malware, the one that we're in. There's a significant amount of growth and new research coming out which we'll — I'm not really even sure how far it will take us. But there's definitely more than three years of growth that I can see.

Joe: Well, I certainly agree. I don't think there's any plateau in sight. It seems like almost every other day you hear about huge advances in technology and new techniques, and what we're seeing in society at large is just breathtaking in terms of the speed of technology. And I guess the growing number of malware creators are going to certainly keep you busy. So, yeah, what you've shared with us so far is really interesting and exciting. I look forward to maybe being able to reconvene and having a similar conversation a little bit further down the road and see what you're up to. So, thanks very much from me for the conversation. It's been great; I really enjoyed it. It's been really interesting. So, Jill, that's all I have for today from my side.

Jill: Great, wonderful. Thank you, Joe, and a very special thanks to Darren Chinen from Malwarebytes for joining us today. It was a pleasure to have you on and sharing your awesome insights with us. And I know I certainly feel much better knowing companies like you are out there protecting us from the bad guys, so great job. And to our listeners today, please be sure to subscribe to our podcast and stay tuned for more episodes coming soon. That is a wrap for now, so thank you for listening to the BMC Run and Reinvent Podcast. Have a fantastic day.