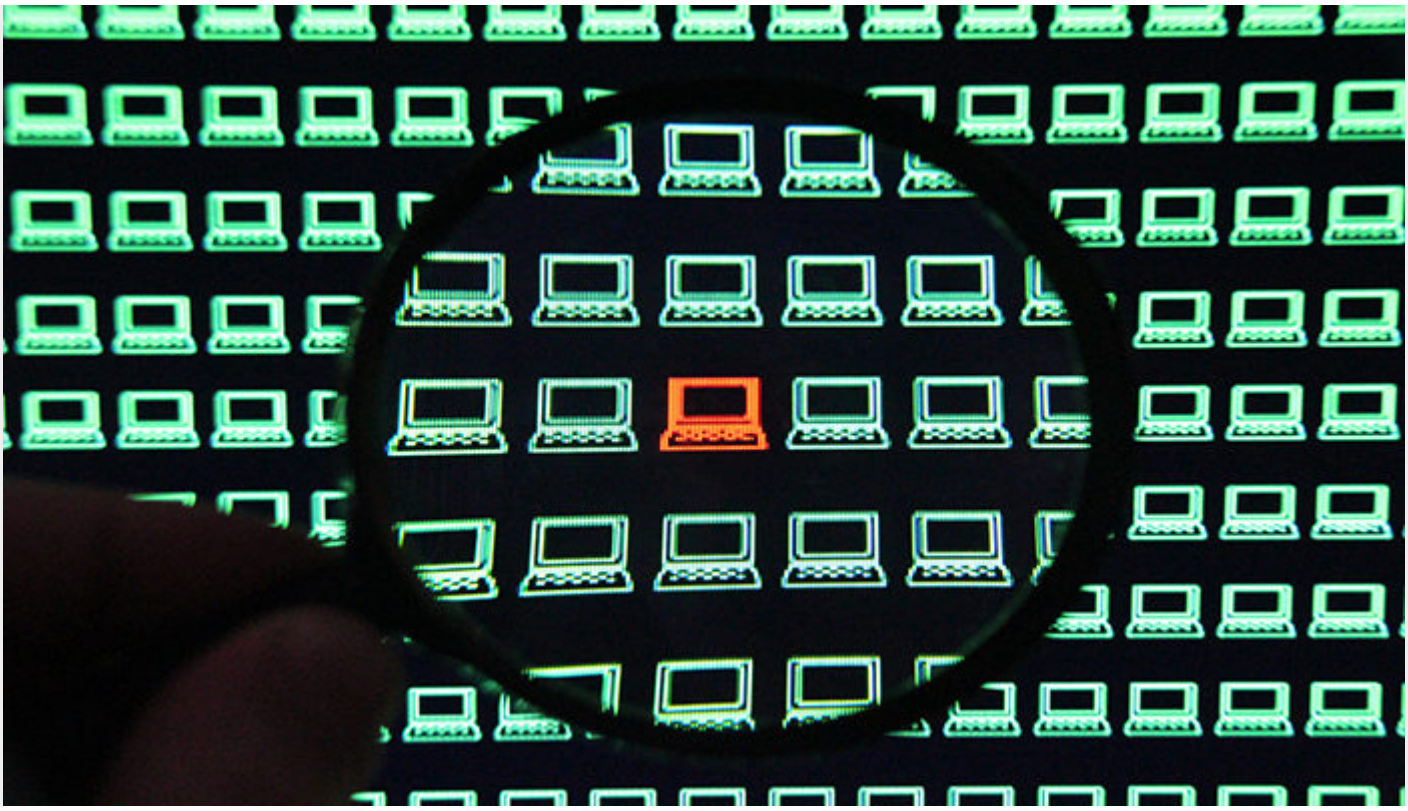
HOW BIG DATA AND AUTOMATION KEEP MALWAREBYTES AHEAD OF THE BAD GUYS



In this Run and Reinvent podcast, I chat with Darren Chinen, senior director of engineering at Malwarebytes, about how big data is helping the company win the war against malware. Below is a condensed transcript of our conversation.

Joe Goldberg: Can you tell us the history of Malwarebytes and how the company was founded?

Darren Chinen: Our CEO and founder, Marcin Kleczynski, 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 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 — our goal is to put ourselves out of business. There's no malware in the world, there's no reason to be in existence anymore.

Joe: Could you maybe give us a little bit of a description about the role of that team within

Malwarebytes?

Darren: I lead 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 IOT data. Think of it as sensors all around the world. And they also do some of the data science and AI as well.

Joe: One of things I think, certainly most people don't draw a connection — certainly not immediately — between malware and malware detection and big data. Can you talk about how Malwarebytes is leveraging the power of the data that you're collecting?

Darren: 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 CI/CD (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: 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: 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: 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.

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 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.

When you walk into your kitchen, you have a machine that is called a toaster, and all it does is 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.

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 DevOps 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 DevOps guy does, I'm going to think, oh my gosh, these guys are infected with malware, or I'm not going to be sensitive enough, and I'm going to think that the engineer and the DevOps 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 very careful about getting good detections, but also not making catastrophic mistakes.

Joe: 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: 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.

Joe: 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: 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? 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: 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: 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 builds. 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.