FEDERAL TECH PODCAST

Ep. 55 Th Role of Synthetic Data in Software Development

00:35
Welcome to the federal tech podcast. My name is John Gilroy and I will be your moderator today. Our guest is Thomas George. He's a senior systems engineer at Midori V i d o r i and yes indeed, we are going to talk about software testing and all aspects of it for the federal government. But first, I would be remiss if I did not tell you that we're recording this from monk's barbecue in lovely downtown Percival Virginia. I think you're eating some barbecue earlier Taos What were you enjoying? By the way?

01:04
I'm a big fan there Brisket. Brisket was quite good. I love their greens, too.

01:10
Yeah, we had a cornbread in there is delicious, great. Unfortunately, we had to go from barbecue to serious stuff, software testing. To tell you the truth. When I was in school, I wasn't a big fan of testing. I hated testing. It was just tedious and boring. And I and he's very good at testing. You know, my wife and daughters are really good at testing. But I'm terrible at it. And when I work with software companies, inevitably they bring in a new guy like Claude, and they say oh Claude will happen to the software testing Oh, yeah, having a bad because they hated it, you know, and it was very tedious. Whoever think today we see systems are getting huge and big and big and big and and hybrid cloud and all kinds of people accessing and new types of data, it really a software testing can make things a lot easier. And that's we're going to focus on today, the whole software development lifecycle and software testing where it fits in. But first, perhaps you can give our listeners maybe a thumbnail sketch of what Midori does.

02:00
Victoria, is principle is used to software test large systems for the civilian systems for the federal government systems grow very big these days. And they have sometimes in excess of over 50 different systems, which makes it a system of systems. And so you can imagine if something changes or some one particular system as a problem in one area, it can be hard to predict the butterfly impacts and effects as it spreads throughout the whole systems of systems. So having a good integration and system test capabilities will allow you to have some measure of confidence that your system will be resilient to unexpected developments.

02:40
When you said butterfly, I immediately focused on RPA robotic process automation and, and you know, a robotic process can work really fast. But if it's designed incorrectly, it's gonna give you all kinds of troubles, isn't it?

02:52
Yeah, it can. Again, it's one of those things where you got to the better you know, your data, and the better you know your processes, the more the more will fit. So if you've got a good set of data that you have analyzed correctly, and you know how it's going to flow, RPA can be something that is very useful for you, I wouldn't recommend it over something that you don't know very much about, or you're not in control of the inputs or things along that lines,

03:21
okay, I'm gonna present a difficult problem, and you're gonna give me a quick solution. So the problem is, we get COVID data coming in NH has got all this data. And what they have to do is have to analyze it. But sometimes they have to make it available to the public, or sometimes they have to make other scientists have taken. And so we have all kinds of problems with with security and privacy with different types of healthcare data. But guess what, there's also financial data, you gotta be careful with a lot of personal identifying information. And by the way, in the military, there's all kinds of information there. And so what what people do is they set up systems in order to test this before they can release it or maybe do some simulations on it. And so the word I'm trying to work is word synthetic data. So what is synthetic data? And what does it roll for testing.

04:06
synthetic data is essentially data that you use to mock up to represent the historic data that your organization has captured and is using to solve some sort of business problem. It's incredibly useful for the cases when you understand the data and you can mimic it up. And therefore if you're using completely fictitious synthetic data, it solves a lot of problems for testing. As an example, you were talking about security earlier, right? So if you don't have to worry about getting in an ATT because your data is CERT is all fictitious, and it's not titled 13 Or it has no PII because you made up the names yourself, then you can you pretty much hat and it still mimics the patterns and has the richness and complexity of historic data that you have, then synthetic data can allow you to stand up your system In your testing environment and test pasture further, another thing is is on the security side that synthetic data can allow you is, say in the awful event that somebody gets access to your test environment that you weren't expecting to. Since you're using synthetic data. At the end of the day, it's all made up information that represents the the richness and complexity, but it's still made up so nothing is hurt or harmed by it. So synthetic data is quite useful from a security perspective, from a disclosure avoidance perspective, I touched on that earlier proof, prevents and protects PII. So I'm a big proponent of using it just alone from those angles. years ago, there's

05:43
a concept called a sandbox, where systems run in a sandbox is this a similar concept to that or the way different

05:49
similar, exact same exact same calm concept, you want to set up your test environment in a sandbox, you know, so you can play it around, take unhappy paths and stuff like that, give yourself the freedom to go in and

run your automated and manual tests with that. And without having the worries of taking down a development environment or something else.

06:10
The last person I did a podcast with had a big contract with NASA. And when I think about NASA, I think they do a lot of simulation. And so this is this is this is simulating datasets. Is that Is that what it's simulating? Excellent? Yes, sir. So is that a digital twin or something different I there's all kinds of concepts, trying to narrow it down. Alright. So

06:29
usually, when people think of simulation, they're trying to think of take a system and model it and have some expected inputs and outputs, synthetic data would be something that a would be the data that is used to go and say, I have X percent of my data is going to have this characteristic or that characteristic, then you would run your simulation against that synthetic data, and just the synthetic data and do it. So in my view, simulation is an action. Whereas synthetic data is kind of a type of noun and maybe some adjectives.

07:05
When you read about what's going on the federal government, there is an organization of CTOs, Chief Data Officers, I mean, never had that five years ago, it's brand new people read about data. And when you look about the quality of data that constantly talking about it, and so what you might be able to do is if someone says, We are requesting a set of healthcare data, and it has personal identifying information on it, and we want to be able to look at it from many different angles, you'll be able to set up some filters and allow a FOIA request to see this information. So it's got to be very carefully controlled, because one little mistake can just get you in a lot of legal trouble.

07:42
Indeed, indeed, you have to, when you're setting up those filters, you will have to make sure that they're complete, you have to make sure that you're not inadvertently exposing somebody through alternative means. And by that I mean, what how can you expose somebody's PII through alternative means? So say, for instance, you have someone in Seattle, and you're just checking the households in a neighborhood, right? And so say some household has a bunch of households have an income of $50,000, or $45,000. And then there's one household with an income of, I don't know, several billion at the end of the day, how likely is it that you have inadvertently released the information through sheer income identification of Mr. Bill Gates, just by doing that, so you have to pay, it's not, it's not just enough for the name, you have to understand the characteristics and the context so that you can protect PII like that,

08:38
I guess the 500 pound elephant in the room is chat GPT and artificial intelligence. And it's apparent that there is a bias in the data that they use to design some of these artificial engines. I mean, there's, I mean, we all know the jokes are what happens and all kinds of strange things happen. But I think people in the federal government, if they're looking at healthcare data, they want to make sure that it's not biased, because if they're gonna look at health care data, they may make decisions that are going to be life and death decisions. And so I

think one advantage of synthetic data is eliminating or minimizing biased or deceptive results, this must be another on the checklist. Ah,

09:15
yes, it is. It's ultimately synthetic data is data that you have controlled and put in the type of noise that you expect it to. So if you want to go and say, hey, I want this kind of distribution of noise, whether it's Gaussian, or uniform, or whatever other distribution, you can control that with a good synthetic data. That's kind of the things that you the trade offs that you have to make when balancing historic data against synthetic data is you have to independently assess if you have a difference between what you expect and what your his history is saying. Sometimes you have to go and say okay, which is the one that I want to build my system around

09:58
on unintended cons. quinces my mind, I raised three kids all through their teenage years. And there's lots of unintended consequences with automobiles in my house. And so I think that's, it's kind of interesting. And so in a, in a perfect world, we'd be able to take a dataset and say, Well, what would happen if we ran it through application a, you can set up synthetic data, you would run through application, you can say, Guess what, it's working great. Or guess what it broke this RPA is robotic process, and we have to fix something. Go back to the drawing boards. Tom,

10:28
in my opinion, that's the blessing of having test, because it's one of those things where people don't see the impact of cars that weren't crashed, you know. So in my, or the headlines that weren't made, because some system was caught early on, some defect was caught before anything happened. And that, to me is a benefit for having strong synthetic datasets. And in a very aware and integrated practice for test for all these large systems of systems that are coming around these days.

11:00
He's a big fancy word back there the words Gaussian, and named after a mathematical model. And if you pick up a book by Nassim to leave the Black Swan, it's like on every other page, and he hates the Gaussian distribution, go back down to the guy who started and go back and forth and forth. It's just a, it's a fantastic book. And so the point is, is that, well, look, you're not worrying about the outliers, because bad things happen. And you have to put that into you. And so I would imagine that this whole concept of outliers and Gaussian distributions has to be built in from a data scientist perspective into synthetic data. Yes,

11:35
indeed, we use a lot of, for lack of better term random number generators that we specify and say we want to have it a frequency of outliers. So if you want to go and say I want a lot of outliers, you pick a different distribution type. Or if you want a normal amount of outliers, you pick the Gaussian, and you use the normal expectation to do things, what your system is expecting. And for the data that you've specifically constructed to be outliers, you want to see how your system performs, whether it's under load testing, does it fail out? Or does it? How does it respond to unexpected situations?

12:14
While you do sound like a data scientist?

12:19
I'll take that as a compliment. Thank you for that.

12:23
Yeah, that's it's kind of fascinating.

12:26
This is Brian veiny. From Aon. Listen to Episode 54 of the federal tech podcast where we explain robotic process automation.

12:33
One concept that's very popular in the federal government now is, is continuous testing. And, yeah, so So that has got to play a role with this whole topic of synthetic data as well, because you could run this scenario couldn't show.

12:49
Yeah. So one of the tricks that is, in my opinion, a great thing is this concept of continuous integration, continuous delivery. And part of that you have automated testing that you can build right in. And so right, right, when a developer gets through coding something and they check it in, and it passes the various tests that the CI CD tools can go in and do, you can use your synthetic data to feed it and see how it performs under this under conditions that are going and you have a control data set to to automatically test as they go in and continuously integrate every time.

13:26
So when I read an article, and it's a shift left, that's enable shift left to do shift left. What are you shifting? Yeah, test sample to see what happens. test scenarios. Well, Thomas, you're a relatively young man, and I'm a relatively old man. In fact, some people think I went to grade school with a Blinken. No, no, no, I didn't go to grade school. I went to high school with cabling. And back in the day, there was a product called Lotus 123. I don't know if you remember that. But that was the predecessor to excel. And I can remember when they released Lotus 123. The advantage was now finally we can play. What if it's the what if scenario? And that what I'm reading about synthetic data? Oh, yeah. Oh, Thomas comes to work. And he was wants to play a lot of what if scenarios, doesn't he? Yes, yes, very much. So what so? Is that the correct analogy or my you off here that you're

14:17

totally on base? That's it. You know, you want to have controlled? what if scenarios that you can evaluate and see is this going to meet my expected parameters? Or is this system falling the way it's supposed to? Is it staying up was supposed to

14:32

there is a well known data scientist in town called David Linthicum. And I know, David, and he wrote an article this week, and he talked about repatriation. Nothing to do with Visa has nothing to do with Mexico or Poland or any other country out there. repatriation is a process where some people, maybe an organization may move to the cloud and then maybe move back and go on prem for a lot of different reasons. So I would think that one approach would be well let's let's have Thomas whip up some synthetic data and pop it in there and see what it breaks. I mean, that's what you're looking at, isn't it?

15:03

Indeed, you know, you migrate to the cloud. So that's a different set of hardware, equipment, software configurations. And we try our best to automate and have everything controlled. But wherever there's people, and there is there room for differences, right. So you could use the synthetic data without migrating your data into the cloud yet, because you may not be prepared for that. So you could use synthetic data to go in and test to see if the stuff that you migrated to is working as the same way as the stuff that you migrated from. Similarly, if you're gonna go back and go back to on prem, see, if you built out your hardware spec, see if you built out all your software configuration in in the same manner that can handle what you were now have gotten used to, from Cloud performance perspective.

15:51

Everyone's walking around with a bias, bias here, bias there. And I think this is one of the concepts about artificial intelligence and robotic process automation that people kind of lead by the wayside. They don't think about it, because according to a recent Gartner report, 60% of data for AI is synthetically generated. Wow. So, so that is pretty big number. And so my question is, you know, are we just gonna see one version of AI and laugh at it another version, laugh at it another chat GPT and laugh at and say, Well, this is a lot more difficult than it sounds.

16:27

Yeah, I think so I think it's going to, it's, uh, to me, it would be a process of iterative improvement. You know, right. Now, let's think about how AI has progressed over the years, you know, let's compare old chatbots to what chat GPT is doing for nowadays and or Google search as an example. So if you were to type in Google search, and you wanted to eliminate you, it's a search on some topic, press A minus, to eliminate those search results right there. You know, and then we iterate on that. And now we can go and say, Okay, I want you to talk to me about this. And then you don't have to use special codes anymore, you can just tell chat GPT, hey, I don't want to talk about this. But please exclude that in plain language, as you know, as an example. And so as we get better and more iterating. Over time, more and more features will make technology easier for people to use, and ultimately be more productive with

17:23

in August, I will be attending this small set conference in Utah. And, you know, there's lots of satellites are being sent in orbit, lots of constellations, 1000s and 1000s in the next few years. And this is, this is going to be a perfect application for synthetic data. So you're sitting in Utah, and you're building a satellite, once you get it up three miles from the earth, you just can't fix it, you have to do is you have to test all the data hundreds of times and make sure to operate once it gets up there. And this has got to be true for all Elon Musk has activities as well. So I think it sounds like an esoteric concept. But no, it's really, really a practical concept to see if the software because he can't change things once it's up there.

18:02
Very true, Huck, although I can't speak to soft satellite development. Yeah, it's a bit it's a bit out of mind. But yeah, I think the concept is there. Certainly, you know, if you get into places were going kind of going back to our security thing. Earlier, if you get into places where it's hard to change or something like that you want to know ahead of time is what is going to happen, is it going with with my data, and if I can mimic the problems that are going to occur, I can at least get an idea of what's going to happen ahead of time. And so I can build in failsafes for the satellites or whatever else. That is there.

18:37
Well times if you're interested, the satellite and satellite technology, there's a podcast called constellations. And they bring in all kinds of people. He talked about software testing and satellite as a service and space as a service. Oh, wow. fascinating concept. Yes, that is, yeah. constellations, podcast, app comm it's pretty. I deal with software developers all the time. I've had arguments about hot sauce with them. I've had arguments about databases with them. I've been called an idiot and several languages. And it's probably true. And there's something they talk called the software development lifecycle. And so let's say we have a whiteboard, and you're eating some hot sauce, and you're, you know, picking on me for anything about barbecue or something. And, and so if you were to graph this on a whiteboard, where does testing fit in the software development lifecycle? It seems like it's circular. Where's it fit in?

19:24
Well, a couple of places. So you could start and say software testing starts with unit testing at that development level, right? And then after it's proceeded along, and you know, they check their code. And ultimately, you want to go and say, before I bring this to some production environment, let's go and have a staging environment or a test environment to see how it does outside of my development environment. And that's another angle that you come in and test test right there. And then you say, Okay, I've got this one system. How does this system talk to other systems, right? And so you have multiple whole systems connected together and your testing or staging environment or whatever. And then you see how they talk to each other. And then another angle that comes in as you go and say, Okay, after we've gotten our blessing from the developer, and we've gotten our blessing from the systems integrator or the systems tester, then let's have the end users come in and see is it is this matching their expectations. And so you have user acceptance testing that comes along at the close to the end. And then they go and say, Yeah, this is meeting what I was hoping it to do. Or maybe we could tweak this or this or that to make it more useful for me. And then after that, the developer has long since been working on the next thing, the tester has been working on the next thing too. And, you know, you just keep that circle going, that you were that you were saying is, the user starts with

developer, it goes to us, ultimately, it falls way to user acceptance testing. And then while the user acceptance testing is going on, the developers are busy developing the next thing and that chain. And you know, feedback comes back from user acceptance or systems integration testing. So it really is a circle and it fits in and all of the all the places that I can think of for it,

21:09
well, something also fits in here. It's a tool that your company has. So if you go to the Dory, it's WW one dot, Midori, V i d. O ri.com. I think you have some kind of a performance testing tool. Is that right?

21:23
Yes. It's the quite imaginatively named Midori performance testing tool. Oh, reginal. Yes, we're engineers. That work. The performance testing tool is designed to help load systems under. And what I mean by loading a system is go in and simulate hundreds of 1000s of people accessing some website at the same time, it's used to go in and say, Can this database take all this information at once without breaking or having unexpected load times? Because nobody wants to go in and have their web page take a couple minutes to

22:02
load anymore? So So actually, it's designed specifically for large datasets? Indeed. So we'll have to tell the people in the satellite community to use this tool when they design a satellite that's going to go to Mars because they can't fix it once. It's like Earth orbit? Yeah, not at all. So we're going to come back to the Dory in the future and talk more about testing. Talk more about synthetic data. Where do you think it's going to head the next three or four years? Just in general, I think more and more people are going to give more credence to testing? Or do you think you're going to it's going to have to be an incident for people realize it's going to actually investing in testing is going to save them money in the long run?

22:35
That's a that's an interesting question. I tend to think that testing is going to I'm going to take the positive view and say that testing is going to become more and more known for what is used as the systems get bigger and larger and more integrated with each other. I think that, in my opinion, is it becomes as essential as having developers and anything else in the area, because of how hard it is to predict the downstream impacts from some butter from some chaos butterfly, upstream. And to me that that is the value add that you need to and the drive for it. Admittedly, it isn't as showy as some of the other things but because at the end of the day, there aren't any newspapers, statistics on car crashes that never happened. Right. Right.

23:20
It's a precaution. That's interesting. We'll have to expand on that in the future book. Thomas. We're running out of time. You've been listening to the federal tech podcast with John Gilroy. I'd like to thank my guest, Thomas George, senior systems engineer at the Dory.

23:37

Thanks for listening to the federal tech podcast. If you liked the federal tech podcast, please support us by giving us a rating and review on Apple podcasts.