On Cloud Podcast
December 2023

The Deloitte On Cloud Podcast

David Linthicum, Managing Director, Chief Cloud Strategy Officer, Deloitte Consulting LLP

Title: **VMware's Alexander Romero shares strategies for implementing and navigating Generative AI**

Description: In this episode, David Linthicum talks with VMware senior director for cross-cloud services, Alexander Romero, about the impact of Generative AI. They discuss issues that companies face with implementing Generative AI such as cost, complexity, information privacy, security, and intellectual property concerns. They also discuss VMware's solution framework for these issues. Finally, Alexander shares his views on Generative AI's future risks and benefits and how companies can navigate their Generative AI future.

Duration: **00:24:24**

David Linthicum:
Welcome back to the On Cloud podcast. Today on the show I'm joined by Alexander Romero. He's a VMware senior director at the VMware cross-cloud Services, obviously focusing on multi-cloud., where there's a big push for VMware right now. How are you doing, Alexander?

Alexander Romero:
I'm doing great. Thank you so much for having me on the show, David. It's always a pleasure to join you on these thoughtful topics and challenges in the IT world.

David Linthicum:
Yeah. I think it's a different way of thinking too, as you're seeing. He does his own show at VMware, which I was on. It was a great show. What's the name of the show so people can look it up, by the way?

Alexander Romero:
It's the Multi-Cloud Expedition. We're up to episode number ten. We just finished episode number ten yesterday, actually, or depending on when this airs, before this show.

David Linthicum:
And he's just knocking it out of the park in terms of listenership, so make sure you look that up, especially if you're into understanding what complex cloud architectures are and different solutions that are out there. And it's a very balanced show that looks at all aspects of it, really kind of looking out for you, the enterprise user of this technology and making sure you get to a good, valuable conclusion to all this. So, tell us the Alexander story. How did you get to VMware? Where did you come from? What have you been focusing on? What did you do recently?

Alexander Romero:
Yeah, wow. Well, it's many years, many years in the IT world. It started out a long time ago, actually. My first job out of college was at Procter & Gamble in their IT department. That was really a fantastic opportunity to understand the inner workings of a very large, huge enterprise company. They had about 150,000 employees at the time and worldwide operations.

So, it gave a sense of the need for scale of operations as well as how those operations must deliver value. Then fast-forward to starting a company out in California that I ran for four years, a stint at Google also, as well as MobileIron – I'm kind of going fast through my history – before winding up here at

VMware about five years ago. Here I'm responsible for all the product and technical marketing of VMware cross-cloud Services. So, think of that as the way our services come together, all of our products and services, to solve the multi-cloud challenges for customers.

David Linthicum:
I think it's interesting also that if you look at the VMware marketing materials and where the market is going, a lot of the industry, a lot of the enterprises I work with are moving to more complex cloud architectures, multi-cloud. You guys are aligning directly to the need, which I think is unique because the hyperscalers are focused on their particular silos and how they think cloud should exist I think the important stuff is the connective tissue that exists between all the service and target systems that are out there. You guys have positioned yourselves in that space, which I think is great. Tell us a little bit about that, that strategy behind that.

Alexander Romero:
Yeah, wow. We are definitely on the same page about that, David. Customers, for the indefinite future, they are going to be in a multi-cloud world. They already have previously a huge on-prem presence, anything on premises. Then cloud came in. There's lots of benefits to the cloud, so a lot of cloud first companies put new capabilities in there. Some of them did lifts and shifts. Then what's kind of emerging is edge. None of these are going to win out. When I say "win out," I mean win out as the ultimate solution for a perfect business outcome with the perfect cost model. Instead, it's a mix between them. But the challenge now is how does one balance that benefit of the complexity – because there's benefits in the cloud.

There's benefits to on-prem. There's benefits to the edge – with the cost and performance measures, and people requirements to maintain what is growing to be more and more complex architecture? So, a little plug here for VMware and cross-cloud Services is really looking at that problem, because this is a problem customers have been coming to us for several years now, and that we have got a number of products that we assemble in order to help them in that complex architecture.

David Linthicum:
Yeah. The great thing I think in your strategy is you're not trying to replace things. You're trying to enable all this stuff to work and play well together and provide core infrastructure to make that happen. Is that accurate?

Alexander Romero:
Yeah. It's not realistic to isolate on a single vendor or a single cloud or everything on-prem. So, the real-world problems that customers are facing is: How do they make all this work together as efficiently as possible to get that business outcome? That's really where we have been focusing, on abstracting that complexity to help them get through that, because it is. There's so many kind of funny stories about this. Customers will come and talk to me and they're like, "Yeah, I'm a swivel," what do they call it, "a swivel armchair, a swivel chair admin," meaning that they swivel between one console to another console to a third console, trying to figure out what's going on in their infrastructure, how they resolve tickets. The list goes on and on, but that's just part of that complexity that they need help with.

David Linthicum:
Yeah. It's going to be important moving forward. Obviously, if they're focused on too many things, you get into breaches, human errors are made, outages occur because we're not paying as much attention to these various systems. If you look at the postmortems on a lot of these outages and a lot of these breaches that I see, it's because there was just too much going on in front of the human beings that had to monitor it. So, definitely moving in that direction. We're able to extract complexity and deal with automation as core to moving forward. Let's talk about a particular problem.

Let's talk about AI, specifically Generative AI. One of the things I was writing about this morning, as well is trying to figure out how to build infrastructure around this stuff, and also how to pay for it. That's two big problems that I think are in front of my clients that I'm getting a lot of questions around that. So, what's changing in terms of us adopting Generative AI, predictive analytics, predictive AI, things like that? What are the enterprises looking at?

Alexander Romero:
There's been kind of a good history of the predictive AI component that does exist in a lot of companies, and much more in larger enterprises. When we think about how to distinguish the two, between predictive AI and Generative AI, predictive AI is going to be more about crunching a bunch of data in order to get to a specific answer, like is this person a credit risk, given all of this history of information, as well as the history of their peers and the geography. It's all that, but it kind of gets to that one answer that companies are looking for. Now with the advent of ChatGPT, we call ChatGPT this kind of iPhone moment for the consumer, but then that plays over directly into the enterprise.

So, folks that are just using ChatGPT for personal kinds of questions and summarizations, et cetera, go, "Wait a minute. This would be great if I could do this at work for…" whether it be IT infrastructure or sales or you name it. So, that has created a lot of interest in how can enterprises use Generative AI capabilities to advance any business function. Because really, this could be applicable from the legal function to the accounting function, to the sales function, to the code function.

This is kind of interesting because there's also a bit of a freak out moment that's happening, which is this stuff is so good and so consumer-friendly that a business had people who were actually taking proprietary code, putting it into the public Generative AI services, in order to get advancements in that code and then put it back in. That ended up being a big no-no. But then this creates this problem where enterprise wants to use this technical capability, but they've got to make sure that they use it in a responsible way that doesn't disseminate their private IP. This is kind of the big challenge that they're figuring out. How do I get that technical capability, but still protect the stuff that's kind of the special sauce?

David Linthicum:
Yeah. I think that's important and we need to look at it moving forward. What about infrastructure class and storage? What are other challenges that the enterprises are considering right now?

Alexander Romero:
Well there's definitely a cost question. How can this be cost-efficient? As well as we go down the line in terms of performance. Performance is another big question. How do you get the most performance out of that money that's being used. Then also privacy. There's the privacy aspect of making sure that whatever Generative AI models, any large language models that are then fine-tuned with proprietary data do not get outside the perimeter. They don't go out in the public domain. So, that's a huge concern. We just talked about that a little bit.

There's another level, which is to construct the right access controls, so that if one is in, again, let's say the supply chain organization, well, they shouldn't be able to go ask, even if it's an internal Generative AI capability, questions about what's the product roadmap, for example, or maybe specifics around IT infrastructure, things that they don't need. So, that's another concern of enterprises, which is, well, if we train these models with our own proprietary data, we've got to make sure that it gives the information back only to the people that should have access to that information.

David Linthicum:
I think that if we're going to use this stuff at scale, all these problems need to be considered. You need to consider that, certainly dealing with proprietary data, certainly even if you understand how data can take on other forms. You could put in anonymized data and actually come up with proprietary data based on the fact that the AI system is able to find patterns that aren't there in the anonymized system, but will be there when it comes through deep analysis and understanding those patterns. It actually will determine what the data means, and therefore put up data that may be PII information that's derived directly from analytical engines.

People are freaked out about that right now, because if you think about it, if you put in data that really does not have protected information, it's completely anonymized, it's anecdotal, and it's able to produce information that becomes more risky and actually is PII information. I don't think people plan for that, yet that's what we're seeing in many instances, also data biases.It's never going to be completely unbiased, but there's some issues around that that we need to look at.

So, how do we solve this problem? What should enterprises be considering now? What are some of the solutions that are starting to arise out there that we need to take a look at? Are we going to have to go through and figure out these problems first, or are we going to have technology that we can depend on to remove some of these problems or at least eliminate part of them?

Alexander Romero:
From a framework standpoint, I think for audience members who are getting requests for AI-capable infrastructure in order to solve business problems, considering these key categories, like five of them, privacy is one. We talked about that. Choice is another one. We haven't talked about that in detail, but that's choice of language models, choice of other components that might improve outcomes. Cost is obviously important. Performance is another one. Then that compliance that you were just talking about, that access control are two different things. So, those five, and I'll go through them again, privacy, choice, cost, performance, and compliance, those are five big bucket items.

Now in a perfect world, those are kind of explored. I think it goes back to even some of the bigger multi-cloud architecture questions that you and I have batted around, which is really kind of begin with the end in mind. So, what's the outcome one wants from –? I always think of this as a technical capability. It's a technical capability that can in turn generate a competitive differentiator for the company. So, in a perfect world, customers are able to look at: What is it I'm going to get out of this investment? And how is it going to benefit the overall bottom line? It's super-hard to do in reality, in this wild west world of AI as well as the infrastructure for it. So, I'd say that a lot of this will look kind of like cloud first, so to speak, meaning people go and jump in with two feet because there's an urgency.

There's an opportunity. The benefits, maybe they're exaggerated a bit. Then there's really not a good understanding of how to cost it, how to get the performance out of it, and the utilization of it. So, I think it's going to be a lot of learning how to fly the airplane while also working on it at the same time.

David Linthicum
So, I think it's going to be touching the stove, I guess, to your point. In other words, it's going to be businesses that may have to do some hopefully minor damage, before they take actions and start making changes in how they deal and manage this information. I see that as well. I think your analogy to going back in the original adoption of cloud is probably apt. People really kind of jumped in with both feet. Some people were never going to go to the cloud.

Some people were doing all cloud, cloud-only strategies, and moved, in essence, too fast and ended up making some core mistakes that they had to undo. If you think about it, some of the repatriation that's going on right now and moving back into private systems is kind of a result of that. In other words, we moved too fast, didn't do enough planning, didn't have enough rigor around it and enough compliance and security, and we got ourselves into a bit of a pickle and now we have to redo some of those things. So, it wasn't completely a waste of time, but some of these systems just didn't bring the value to the business that they thought they would bring.

I think this is exactly the same sort of scenario that's playing out, probably a bit faster just because we already have the cloud stuff in place. We can provision this stuff on demand in a very short period of time. So, we're able to, I think, get to some of these issues that you just mentioned quicker and with a lot of forethought and planning. So, we're moving at the speed of need, but I think there's going to be some things we do that are going to be mistakes. But I think in many instances we're going to have to make those mistakes and learn from the mistakes. I think if we're too cautious, that could be another problem as well.

So, what solutions does VMware have in this space? What are enterprises considering now? What should they consider in the future? Talk about what your customers are seeing in this space, and how you guys are applying these kinds of solutions to solve the problems.

Alexander Romero:

It's that kind of framework of things that I mentioned, with the focus being on the privacy of IP information. One of the things that VMware announced most recently, at our Explore conference in Las Vegas, was private AI and the private AI framework, which allows for a reference architecture, which I'm happy to drop in a link or have it even kind of support that. You can also go search like VMware Private AI. But it's going to provide that framework in order to put together the infrastructure to enable the data scientists to get their work done.

The simple analogy I always use is, okay, a data scientist is working on something on their laptop. They have decided that they need to train this model. Their laptop isn't insufficient, or their desktop capacity-wise, to do so. So, they open up a ticket, say to infrastructure, "Hey, I need capacity to run this model." Then it's handed off and this is where VMware comes in. It actually comes in together with Nvidia. Nvidia and VMware have been virtualizing GPUs for ten years now, which is a fairly long time. So, most recently, in the last three years, tremendous advances in virtualizing it specifically for machine learning and AI, but that GPU virtualization has been going on for a long time. So, for the admin that knows the VMware infrastructure and how to allocate it, they can allocate specific GPUs to somebody that needs it in their organization in order to run that higher capacity type of workload.

So, there's an approach, which is the VMware private AI, but then together with Nvidia then we actually will have a single type of SKU, and this is coming later next year.If the line of business is asking for infrastructure, private infrastructure that can in turn be used to calculate and train models, because it combines not only the Lenovos of the world and the HPEs of the world, but the Nvidia plus the software layer from VMware to manage and run that actual AI infrastructure. So, this is where VMware can help out, is in that lifecycle management, we'd say, of that workload in a virtualized manner.

David Linthicum:
Walk me through an example of that. so I'm a customer. I come to you. I have these issues. How do I leverage the technology, the stack that's referenced in a way that's going to help me move closer to a solution?

Alexander Romero:
The first thing that comes up is being able to provide the infrastructure, in many ways with the skills that those VI admins already have. So, the same way they would take and spin up a virtual machine, well, they can spin up that virtual server and then allocate specific GPU capability to it. So, I guess that's the nuts and bolts of how they would do it, but it would start with a line of business coming to the IT and saying, "We need this many GPUs of capability." Then talking to VMware, be able to buy that in coordination with an HPE as well as like an Nvidia, and that actual physical hardware would go into their datacenter, in this case, and then from there being able to allocate that capacity out while keeping that privacy layer, so that whatever data it's trained on doesn't go outside of the perimeter.

It doesn't become part of the public domain. Although it's early days right now, so I'm talking a lot about on-premises stuff, this is primarily because the majority of proprietary data at large, in the enterprise, is still – a lot of it is on-premises. So, there's a reluctance from customers, rightfully so, to go and do vendor lock-in in the cloud or transfer their data and worry about egress charges. So, all of that is still being kind of figured out from what's the most cost-effective thing, and also what would protect their IP. These are areas where we can help provide that infrastructure and that software layer to manage it, even as we look to the future and where there might be different clouds that are used for different learning models or for different infrastructure and cost even. So, it's still the wild west, David, in terms of how these things come up, how they're used, and how they're costed.

David Linthicum:
Yeah. I love the solution and the fact that you're leveraging the information where it exists, and you're considering the security parameters around it. This is not, "Hey, you have to move to our technology to make it happen, and make these very risky and very costly changes that are occurring." You're looking to leverage something that I think is going to be in many instances the optimized technology that they should be leveraging. So, let's switch gears. Let's go in a time machine, five years. We're doing this podcast again. Hopefully we're still alive. What are we talking about, do you think?

Alexander Romero:
I definitely think that the technical capability of – I use those words often, which is cloud computing is a technical capability. So, Generative AI is a technical capability. So, five years from now, I think that it will be equalized as a competitive differentiator in certain functions of the enterprise. Let me pick on one, and again, this is a bit of a prediction, but let's say customer service, being able to do customer service.

That I think will mature quite a bit in terms of being able to use Generative AI to improve customer service outcomes and probably will be fairly equalized. Anybody that doesn't do it will probably be kind of left in the dust. The more fascinating things would probably be around code generation, how this code generation is impacted, and also – boy, like five years from now – probably what type of workforce will be needed and where the specialties will lie, whether it be in prompt engineering or how Generative AI will be improving, creating new business differentiators.

David Linthicum:
I think that's pretty much where it's going to be. One of the things I'm kind of taken aback at, when we go forward and look back to where we came from, how things have progressed at a much slower pace than we thought they would, because of budgetary issues and the ability to adopt the technology. Also, you mentioned this at the beginning of the podcast, the ability to go through learning cycles and understand things better, and do some failures and come back and do some successes upon what we learned from the failures. So, where can we find out more about your show, you, VMware on the Web?

Alexander Romero:
We can drop some links if that's supported in the podcast or in the show notes or something. You can just search for Multi-Cloud Expedition, and from there you'll be able to see the ten episodes. There is a range of episodes, everything from security and networking all the way through to private AI, which we were talking about today. So, lots of great material and many different subject matter experts, including yourself, have joined the show and provided that view into it. I had one more thought as we were talking about that five years from now.

I don't think this will be solved, but if it did with Generative AI, it would be transformational. So, here's my thought. Every customer has got technical debt of legacy applications. If there is a way for Generative AI to eliminate that technical debt, that would be truly transformational. I've never seen it. So, when you asked the question, what will we see in five years, one of my first reactions is, well, in many ways enterprise will hang on to the legacy technologies they've had for years and years and years, because they're just to darned expensive to move off of or to refactor. I don't know. If they can solve that, then you'll see tremendous advancement, innovation, and transformation.

David Linthicum:
Yeah, I think so. That's good to aim for, because I think that the ability to leverage this AI capability, which is going to grow stronger and stronger and stronger to actually solve real problems, some of these issues inclusive of technical debt, is going to be an outstanding feature for that to have. Anyway, if you enjoyed this podcast, make sure to like us, rate us, and subscribe. You can also check out our past episodes, including those hosted by my good friend, Mike Kavis. Find out more at DeloitteCloudPodcast.com, all one word. If you'd like to contact me directly, you can e-mail me at dlinthicum@deloitte.com. So, until next time, best of luck with your cloud journey. Everybody stay safe. Cheers.

Operator:
This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to Deloitte.com/about.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Visit the On Cloud library
www.deloitte.com/us/cloud-podcast

About Deloitte

As used in this podcast, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.
Please see www.deloitte.com/about to learn more about our global network of member firms.