Deloitte.

The Deloitte On Cloud Podcast

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

Title: **Generative AI is the future. Now's the time to harness its disruptive power.**

Description: Everyone seems to be jumping on the generative AI bandwagon, and with good reason. It promises to disrupt nearly everything we do and every interaction with technology. In this episode, David Linthicum talks with cloud architect and consultant Lynn Langit about generative AI and how it can potentially play a significant role in our daily lives—especially for businesses. They also discuss how to harness the full potential of AI, and they debate the future of AI and quantum computing.

Duration: **00:25:45**

David Linthicum:
Welcome back to the On Cloud podcast. Today on the show I am joined by my good friend, Lynn Langit. So, welcome to the show, Lynn.

Lynn Langit:
Hello.

David Linthicum:
We know each other. We're both LinkedIn Learning instructors and probably have been running in the same circles for several years. Give the listeners an introduction of yourself, how you got here, what your background was, and what you've been doing lately.

Lynn Langit:
I'm an independent cloud architect. I've run my own boutique consultancy for 13 years now and I learn/build/teach. Maybe not in that order, maybe build/learn/teach. How you and I are connected is through are many, many courses on LinkedIn Learning. We kind of share a student audience since we both are teaching on cloud topics. In addition to that, I've done a lot of work on migration to the cloud for various verticals. Most recently, I've been working with bioinformatics.

David Linthicum:
So, you work with different companies, doing consulting and building something. You also teach things. Build, teach. What was the other one, learn?

Lynn Langit:
Learn, yeah.

David Linthicum:
Got it, yeah. How is it going?

Lynn Langit:
We're in a constantly evolving, maybe more rapid, cycle than normal these days with all the new stuff coming out.

David Linthicum:
Yeah. It's fast and furious. What part of that do you like most? Do you like building, learning, or teaching?

Lynn Langit:
Well, I like having a balance. If I spend too much time learning, then I'm going to be poor because I'm not getting paid for that. If I spend too much time building, then I get kind of stuck in the details of the day-to-day. It's important to know that, but I also want to look forward, and if I spend too much time teaching, then I'm too much looking forward. So, I really have tried to strike a balance throughout my consultancy.

David Linthicum:
Yeah. I find the same thing. You really can't do the teaching without the building and the learning. So, this thing where you can operate in a vacuum and just become a teacher and just create content, that's never going to be the case. Also, it's not only learning, but having pragmatic applications of what you've learned, so the ability not only to learn architectural techniques, but actually work on a project and build things. I've been a builder of technology for 30 years, and I always go to that as the reason why I teach, because you're able to instruct people on how to do something purely through your experiences.

Certainly, you're learning things, which provides a foundational learning aspect of the stuff that people really are looking for. What's the inside base for this stuff? How do I build this thing? What's important for me to look at? What are the top five things I need to consider? Really, get to this pragmatic application of this technology fairly quickly. I think that's what people are desiring to understand.

Lynn Langit:
Yeah. The other thing about teaching, especially teaching from a platform, is you grow a worldwide audience, and, so, you build a community. When you can—you can't always answer questions, but when you start to see commonality in questions, which point to gaps around vendor offerings or around useability or around market needs. So, learning has the benefit of learning on your own, but teaching also provides a learning benefit when you teach at scale, in my experience.

David Linthicum:
Yeah, it does. It also allows you to refine your thoughts. When I write articles and books, it's not only that, but the ability to kind of organize a concept in some sort of a logical way that has more meaning for yourself, so you can get that and, therefore, explain it to people a bit better. It's one thing to understand something, another thing entirely to figure out how to explain it to somebody, so they can get the same level of understanding and, also mixing this up with the pragmatic stuff.

We have to build stuff. We have to deploy stuff. How you go off and do that I think is incredibly important as well, so the trifecta. How are things going to change as we're going through this revolution or re-revolution in generative AI, in terms of teaching content and our ability to understand? And how does it relate to cloud? Are the cloud providers the center of the universe, where we're going to get this generative AI technology from? They're going to drive different engines to build content automatically. What changes in our world?

Lynn Langit:
Well, the first change is: can I accelerate myself? As a consultant, I'm constantly building POCs across different vendor platforms, different languages. It's one of the reasons I allocate time to learning. Can I adjust that time? How am I going to learn differently? How am I going to use the popular tools out of the box? Am I going to create custom tools? It's kind of a whole new world around learning tooling that I'm starting kind of with my own environment and subcontractors that I work with. What would I expect of my subcontractors in terms of learning?

David Linthicum:
Yeah. Also, where are we doing the redundant work? In other words, where are we working, where we can do things and have it automated through this technology where it's not currently automated? I've taken AI-generated courses before with a simulated voice response, and it's not that bad. So, you get very close to understanding and defining how to build these courses. Of course, they're dealing now with a certain amount of information, but eventually, the amount of data that they're going to be able to accumulate, the models they're going to be able to train, they're going to be at a point where we're typically going to need to reference those things more often than referencing our own brain. Am I being overly skeptical?

Lynn Langit:
No, I don't think so. It's interesting because another way that I look at it is addressing skill gaps for my customers. Lots of times they hire me because they have skill gaps in their teams. Typically, they have classroom instruction or customized instruction. Because I have been using some of the tools as a beta user now for over a year, I find myself being annoyed, for example, when I open a code IDE and there isn't some sort of gen-AI integrated. It just seems sort of backwards once you start using this set of tools. So, it really is a spectrum of usage.

Another area that I find myself using more and more is I have to get domain expertise in bioinformatics, which is not trivial. So, the summarization of published papers has been a godsend for me, because me, pre-gen-AI, like looking up terms and trying to figure things out just was so time-consuming. So, there's just a number of different planes, across which I'm getting productivity gains, and I'm really wanting to apply that to my customers and to my students. I'm trying to figure out how to do it.

David Linthicum:
Yeah. I think we're going through that evolution now and I think it's changing the game for lots of things. Do you think the cloud providers themselves are going to be able to leverage this technology to create any more skills? Certainly for the more—detailed skills. Instead of doing cloud architecture and understanding the fundamentals of cloud storage, we're actually looking at a particular storage service and how that works and the API-level compares and things like that. You get down into this tedious level of detail that people need to know to make these things effective. Do you think they're going to leverage this technology to build more skills out there? Of course, there's a going to be a lot of them to sell more services, but will also be able to create some of the talent that I think we need right now in the marketplace.

Lynn Langit:
Yeah, not only that, but, even more fundamentally, gaps that customers still aren't addressing such as infrastructure-as-code, something as simple as using incorporated gen-AI in IDEs to make that possible for some of their teams that maybe are coming out of legacy systems or non-cloud, not only application development in modern languages, but all the surrounding parts and pieces. Another area I think that is going to be quite impactful is improvement of security in cloud, because that's been the ongoing problem since cloud was launched, that there's not been enough security professionals. So, I really see this introduction into IDEs is something that I'm advising my customers to pay attention to, in addition to the large language models. I think there is a use case for both of them as customers are optimizing their particular cloud workloads.

David Linthicum:
Yeah. I think it is going to be game-changing moving forward. You're right. I deal with and interact with various system. I am frustrated too when I don't see some AI engine that's helping understand how we're doing the coding. You can define a baseline application via an AI system. You just mentioned infrastructure-as-code. I think one of the opportunities here now, people are moving into the serverless world. The whole thing about serverless is it allocates just the amount of resources you need to run a particular function, and then it returns those resources back to the pool.

In some ways that's efficient, in some ways it's not efficient, but your ability to define infrastructure-as-code through some "automagic" generative AI system that's bound to your application code could be a more pragmatic way to do it and get more value out of that technology. Do you think we're seeing those sorts of moves moving forward? People are looking to leverage the technology to, in essence, make it cheaper to run cloud services because we're running in more intelligent and more proactive ways?

Lynn Langit:
Yeah, I would agree. I think there is definitely movement towards that. In addition to that we have this whole group of foundational models that launched across the major cloud vendors just recently here, which I would really consider a new category of cloud services. The question is: can you use what's out of the box? Or are you going to then, in my case, guide your customers to take some of these foundational models and train them with their own, let's say, application code data, so that something like onboarding a new developer can be accelerated?

David Linthicum:
Yeah. You mentioned, that cloud providers have come out with 50 base models. That's amazing, the amount of production that has occurred in a very short period of time. Do you think that's going to even accelerate more?

Lynn Langit:
I do, but again, for my own situation and contractors I'm working with, there is a lot of hype around this, of course, because there's a lot of potential revenue. The reality is we all first need to learn how to use AI, when it works, when it doesn't, when it's hallucinating, when the model is a fit for a particular case, before we can go and customize and use these foundational models. So, it's fantastic that—

David Linthicum:
Define hallucinating for our listeners.

Lynn Langit:
Sure. Hallucinate is becoming a standard word, when the large language models simply return incorrect information. It could be factually incorrect. It could be made up. It's because the large language models are just that. They predict the next word. They are not based on any ground truth.

David Linthicum:
Yeah. It is funny getting erroneous information back, but again, it's garbage in/garbage out. These things are only—when I explain generative AI, it's reflective of us. In other words, we're putting out the information. It's returning the same information to us and different organizations in different ways, in different ways in which we want to consume it and, in many instances, more logical forms. It takes the information and writes an article about it, by doing different AI tricks around making that happen. When you get into what this stuff is, it's ongoing learning that really is the power that's going to be there.

It's the ability to leverage these LLMs so they can be more productive for us, the ability to look at the different vendors and application servers. I think you kind of hit the nail on the head. You have to be able to know how to ask the questions. So, it's not that the AI technology is not fantastic, absolutely it is, but we don't know as humans how to use it yet. So, I don't think we're getting to the potential of what this technology can do when we're probably missing some of the deficits that are already there. So, what's your advice to someone who is looking to get into this world?

Lynn Langit:
Well, I have a little bit of an unfair advantage because I'm actually trained as a linguist. I speak a number of human languages. I feel like my time has finally come in the world of cloud application development, because I really see English and these LLMs are coming out first with English and then adding other languages, but English, rather than some other types of languages, being now the ubiquitous cloud language as these tools mature, which is really a game-changer. But English with an asterisk, English based on how the model was trained and how the model is designed to work. So, it's sort of like learning dialects, and I can make the example across the United States because probably many of the listeners can relate to this. English that's spoken in the southern part of the United States is very different than English that is spoken in the northern part of the United States. It's the same with LLMs because they're language-based.

David Linthicum:
Do you think this is going to be about a better of understanding of what a language is, such as English, that it sits in the cloud, and our ability to have derivative models of different dialects off of it? I think some of that stuff is kind of operating there, but even then to take it to the next level and get into the biases and behaviors of the particular demographics that it's communicating with, with the language, and different ways to motivate people. What I'm getting into is applications for this stuff could be the ability to understand.

We have seen recommendation engines for a long period of time, who you're dealing with and the ability to communicate with this in more customized ways, so they can have better learning experiences or, for businesses, better sales experiences. They can increase their revenue by being more innovative with how we're communicating with the outside world by using these base models, and then building these different derivatives on top of it and have this automated, proactive learning in between these various systems, where we can improve incrementally as we move forward.

Lynn Langit:
Yeah. As an instructor, I'm trying to figure out how I can guide my students to use these tools, and how I can incorporate that in my courses. I certainly haven't solved for it, but it's a question on my mind. So, which of the LLMs are going to be most seamless to use? Which are going to have more friction? Some experiments I have done are based on which cloud vendor the LLM comes from. What is the quality of programming language support? Because certain cloud vendors have certain biases towards different programming languages, and I have found that the tendency to hallucinate in the less used programming languages in the particular LLM is something to be aware of and useful.

David Linthicum:
In other words, becoming unproductive out of trying to be more productive because you're getting bad responses back.

Lynn Langit:
In certain languages. So, LLM A from vendor A is great at language A, but not language B. LLM B from vendor B is great at language B, but not language A.

David Linthicum:
So, how can we prepare? Our listeners are thinking in terms of we know the generative AI stuff is there. People are starting to figure it out. They've seen the wonderful things it's able to do because we're able to write thank-you notes, where thank-you notes did not exist. No one knew how to write a thank you note, but evidently we can do so now. But your ability to, in essence, find an application for this, what should they be looking at right now? Should they be looking at different cloud providers? What aspects of the cloud providers would they be offering or developing in the market, that should be something that people who are newbies in the world of generative AI systems should be looking at?

Lynn Langit:
So, if you don't think about the generative, you just think about the AI, so you can familiarize what you do. The results are different than what you get in the search engines. So, you start doing comparative to see. Then once you start to understand that level, as I mentioned, you want to work in your IDE. So, whatever incorporated AI tool, you want to start using some of those.

After you've done that, then you want to crack open these new 50 base models and figure out if you have text-to-text, text-to-image. What's your use case? Then start looking at using those out of the box, because you may be surprised that out of the box it might provide the value, so you don't have to go through the effort of retraining or transfer learning or all the other ways you can augment those models. They might just work for you. There are starting to be efforts now in the open-source community to evaluate the models in terms of how they perform on certain tasks and they are evaluated from other models, and they're evaluated from humans and scored. So, it's this transparency into the quality of the models that I think is going to help their usability.

David Linthicum:
Yeah, and even integration of the models and the ability to have different layers and different ways in which we derive knowledge from existing knowledge, things like that. It's just a fascinating world. It's kind of amazing that we're this far into technology, but right now, just having this conversation with you, I

feel that we have so much more that we can do. Every time that we just peel back the onion, there's lots of different technologies that we need to leverage and move in different directions and things we need to understand. So, what do you think are going to be the killer applications for generative AI technology in the cloud say for the next five years? What's going to be that one application or two applications that businesses are able to apply, and that really kind of change the game in terms of leveraging this technology to a more productive end?

Lynn Langit:
I do think infrastructure/security-as-code, because as long as I've been working in cloud, security and infrastructure automation have been problematic and have cost businesses undue pain and revenue when set up incorrectly. So, the automation of those sort of mundane areas I think has the potential to really impact the value that customers get out of cloud.

David Linthicum:
Yeah, and that's going to get worse as the deployments out there become more complex with multi cloud stuff, pervasive cloud deployments, things like that, and the ability to just not have an overreaching infrastructure in terms of how we're monitoring and managing this infrastructure. So, that's why we're talking about metacloud and supercloud. All it is, is the ability to deal with these very complex cloud deployments through layers of abstraction and automation.

And, by the way, if you're able to weaponize AI to sit on top of that to make some decisions, and we certainly have AIOps and some other technologies that are working on that space, then suddenly we're able to increase deficiency and optimization by 50 percent versus what we're seeing today, because everybody is leaving instances on and not shutting things down, and spinning up too much memory and storage for particular applications. There's just so much waste out there in terms of how we're leveraging these resources that we've got to figure out some sort of an intelligent way to automate these various systems. What are your final words on this?

Lynn Langit:
I'm always so much in the details because of the nature of my work. I'm kind of being a smart aleck here, but maybe people can finally use Kubernetes *[laughs]*.

David Linthicum:
Yeah. It becomes an architectural religion, so people will move forward and use it in smarter ways. Hopefully, it's that and all other types of technologies. We talked about serverless, but the ability to use Kubernetes and some of the cloud native architectures that are starting to emerge, use them in better and more intelligent ways, so we're not overapplying this technology; and while we're doing so, we're doing so with the optimized amount of resources. I hope that happens.

Let's talk about quantum computing and AI. This is a huge area that we seem to be working at. Quantum computing is something that I've been looking at for a long time. We have the ability to weaponize quantum computing into a space, which is going to make another emerging space, in this case AI, much more effective. What are your thoughts on that?

Lynn Langit:
I have been dabbling in quantum computing back for many years, since I've gotten access to D-Wave way back when. I bumped into those folks at the Vancouver Airport and got on the beta many, many years ago. It's a nontrivial nut to crack for somebody who doesn't have the educational background, but the first point I'll make is if a relatively normal architect developer can make progress, then the rest of the world can, too, particularly with the cloud vendors now offering quantum computing services. Not all of them, but most of them are.

It's really starting to become relevant for specialty use cases. Now I happen to work in biomedical imaging, which is one of the use cases, and there are starting to be papers published on QCNNs, Quantum Convolutional Neural Networks, that show both acceleration in time and quality. So, how that ties into AI is if you can compute on all the things faster, you can test your models. You can optimize your models. You can accelerate AI. So, I see a convergence of those two things as one of the key areas to pay attention to in terms of the adoption of AI in the cloud.

David Linthicum:
I couldn't agree more. I think the cloud services having quantum computing capabilities and quantum computing services is going to make this stuff affordable finally. I mean quantum computing has been around for a long period of time, but who could afford it? I was doing AI back when I first got out of college, but no one could afford to build those systems at the time. It was $30 million just to do a simple application. But this kind of makes it on-demand as a utility. It makes it kind of affordable.

The point you just experienced in leveraging the beta, people can experience themselves for a relatively low amount of money and also find some killer use cases in there. You had the image analysis use case, but the ability to do oil exploration and all these other things that are computationally intensive, the ability to do predictive analysis and these sorts of things that may take three to four weeks to run if we run it on normal infrastructure. But finally, the ability not only to do this, but make it affordable. Is that something you still see in the near future?

Lynn Langit:
The price of the specialty hardware needs to come down, but looking at what the need to train gen-AI has done to the GPU war, I think customers are going to benefit. Of course there are additional types of specialty hardware. If you talk about quantum, of course it's QPUs.

David Linthicum:
Can you define GPUs and QPUs for the listeners please?

Lynn Langit:

Sure. GPU is a Graphics Processing Unit or chip. It's an alternative to a CPU. It allows you to offload certain types of computation that are very related to model training for AI, so linear algebra operations, to get kind of nerdy. QPU is a Quantum Processing Unit, so you have available qubits rather than bits. It allows you to compute on—think of a sphere, basically, rather than an up or down switch. So, you have the possibility of computing all the points of a sphere, which is oversimplified, but it helps to give the listeners an idea of when I say all-by-all-by-all, the computational possibilities that QPUs provide.

David Linthicum:
Yeah. This is where things are moving to. I think this is exciting. I think we have a tendency to overapply this stuff. I think some of the battles I'm going to get into are, "I don't think you should build that business application using quantum computing." However, if the business application needs to do advanced AI and predictive analytics, then there may be some advantage in making that happen. So, we have these use cases to look at in terms of how we're putting the optimal amount of technology to solve the particular problems. I find in the AI space, people have a tendency to overestimate and kind of manage by magazine in terms of how they're leveraging the technology. So, do you think this is going to be a lot of a balancing act and making sure that we don't do things with this technology that shouldn't be done, very much like when we were dealing with AI 30 years ago?

Lynn Langit:
Yeah. The thing that I see on the horizon already, because the cloud providers are already there, so they're sharing this with the customers, which is to use these foundational models and to do transfer learning. I haven't found a customer, and maybe I'm just in the wrong crowd, but who is ready to do that yet. That's why in this podcast I was talking about my own process. This is a new paradigm in many ways and you have to walk before you run. So, you have to understand what these models do and what they don't do before you go and modify an existing foundational model or you're just going to waste your time.

David Linthicum:
Yeah. Keep that mind, overuse of the technology and applying it in the wrong spaces are going to be something we're going to have to deal with moving forward. We always do that when we move to any sort of technology, but I think specifically in this area we can make some huge mistakes in terms of overapplying the technology in areas that it really doesn't fit. So, where can people find more information about you on the Web?

Lynn Langit:
I am a prolific publisher to GitHub. I make a lot of samples for various customers. I will publish, customer information removed of course, on my GitHub. I also, obviously, have many courses on LinkedIn Learning. I write on Medium as well.

David Linthicum:
Great. I'll tell you what. Lynn is the real deal. She knows her stuff. She has got a huge amount of work and experience out there that proves that she knows her stuff, just a huge amount of content, a huge amount of learning, and a huge body of knowledge and work and that's extremely impressive. So, make sure you look her up because she's one of those people that you want to keep an eye on and keep following. So, 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. Let me try that again. 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. You guys 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.