The Deloitte On Cloud Podcast

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

Title: **David Linthicum and Mike Kavis look ahead at cloud for 2024: trends, tech, and insights**

Description: In this special LinkedIn Live edition of the podcast Deloitte's David Linthicum and Mike Kavis forecast all the big trends and tech for cloud, 2024. They predict that 2024 is going to be the year that Generative AI begins to return value to the business, and that cloud optimization and multi-cloud management will become more critical than ever—partially because of Generative AI. They also engage in an animated discussion of how to balance workloads between on-premises and cloud.

Duration: **00:25:35**

David Linthicum:
Hey, everybody welcome to our LinkedIn Live presentation. This is going to be 2024 cloud computing predictions. What's coming up this year in cloud computing? What to look out for? And just putting some lines in the sand as to where we think this technology is going and providing some commentary around that.

So, with me is my partner in crime, Mike Kavis. Mike, why don't you introduce yourself?

Mike Kavis:
Hey. Mike Kavis, managing director at Deloitte. Do a lot of these types of things with my buddy, Dave. And we go way back. And so it was a thing. And it's great to be here to talk about the New Year.

David Linthicum:
Yeah, Mike and I worked at what? Four different companies across four different companies—different relationships? Yeah, it's amazing.

Mike Kavis:
Yeah.

David Linthicum:
Yeah, don't make anybody mad out there. You're going to be working with them for a long time.

Mike Kavis:
[LAUGHS]

David Linthicum:
Anyway, I'm Dave Linthicum, chief cloud strategy officer, author speaker, BLS geek, love talking about cloud computing. And so let's jump right in.

So, Mike, in looking at the predictions this year, there's this technology called generative AI. Have you heard about it?

Mike Kavis:
Yeah, a couple times. Yeah.

David Linthicum:

So, obviously, that was on top of the list because it's on everybody's mind right now. And I'm just talking to the press about it and lots of presentations, a lot of writing, a lot of speaking about the impact of generative AI on cloud computing. So, I have some thoughts here but I'd love to get your thoughts first. What do you think is going to be the impact, at least, in this year 2024 in terms of real projects getting done, real implementations being done?

Mike Kavis:
Well, I have two answers to that, one where I think long term this is going and one where I think going to get done this year. So, if history repeats itself, every new technology spins a lot in the first year and a lot of proof of concepts get done but very little traction from a production world a lot because it's new and it doesn't have all the enterprise-wide security bells and whistles. So, there's always all kinds of resistance and issues trying to unleash or harness the capabilities of gen AI.

So, I think there's going to be a lot of spin in within an organization. Everybody's messing with it, for sure. I think this is the year of trial and error and discovery on it. But I think long term, it really changes the way, especially from development, how we build things, how we produce products, how we abstract things so that non-technical people can produce things as well.

So, I think it's going to be pretty impressive. Obviously, it's only as good as the implementation that you do, but I really a rusty developer like myself can go build an app to my specs, to the way I code. I tell the chat, “Hey, this is how I want to see things. I want to see chest harnesses. I want to see this”. And then I said, “OK, I want to build X, Y, Z.” Boom, it spits it out. OK, well, I've never used Go. So, let me do it in Go. Boom, it spits it out. Now obviously, I have to go fix things, but it's a heck of a lot faster than doing Google searches all day long, trying to remember how to do stuff I used to know how to do. So, it's pretty impressive.

David Linthicum:
Yeah, I think it's going to be a game-changer. I think we just don't know what games it's going to change at this point. Everybody looks at the technology and they go, this is important stuff. It can generate net new content, unsupervised learning, a lot of stuff that AI has never really had to—had in it to make it productive.

I mean, AI has been around for a long time. I mean I'm 62 years old. I was an AI developer when I was 18, writing Lisp and M1 programs. And it just never inflected because the capabilities weren't there to really drive a lot of net new value. Now with the generative AI models that we have, we have just a huge potential to turn businesses into automated smart entities that can provide better customer experiences and better supply chain integration.
But I think the computing public is like a deer in the headlights right now. They don't know what to expect from it. All the cloud computing conferences, we saw this last year, have become generative AI conferences. Everybody's talking about it. Some major acquisitions are occurring. I think the big cloud providers are trying to figure out what's going to happen with it.

And I think 2024 is going to be stepping in the direction of trying to figure things out. We know it's going to be expensive. We know it's a data-oriented problem. I was doing some research into the architectural differences with generative AI stuff because obviously, we're getting a lot of questions around that. At the end of the day, it's really just a data problem. You're dealing with distributed databases that are all over the place. In many instances, we're going to be consuming training data from existing on premise systems. We have to figure out how to do that better than we have right now, have to deal with the security aspects of that.

And to the point you made earlier, which I think was spot on, we haven't figured out the dimensions around that. How it's going to impact the network? How it's going to impact security? How it's going to impact our existing cloud infrastructure? Most of these applications are going to be heterogeneous, even though they may be concentrated in the cloud. And the ability to use all kinds of different technologies that have to be brought to bear to make this technology valuable.

Also the ability to pay for it. When you price these things out, they're about three times the price in terms of the cost of the cloud resources you have to leverage and the cost of the hardware you have to buy if you're running it on your own stuff. So, that has to be dealt with as well, how do we optimize those technologies.

And I think that right now, it's just a learning year around generative AI if you look at what's really happening. A lot of people are writing about in the press and speaking about. It's like there's a big party going on, we're not invited.

Besides some proof of concepts, which you mentioned, I'm not seeing the huge amount of deployments that are done yet. I think we'll start seeing those through the end of the year. What are your thoughts on this?

Mike Kavis:
I agree. As you were talking, I was thinking back to when containers were starting to become a thing. It's a similar story. Like you said, containers have been around forever. But what Docker did was made it very user-friendly for developers to consume. And that's the same thing that's happening here. AI has been around forever. Machine Learning has been around forever. But now it's at a place where we can use it easily and can really change the game.

But I agree. I mean, a lot of enterprise tendency is to do this themselves. The amount of infrastructure required to do the modeling and all this stuff. I mean, I don't know how anyone can come up with an ROI to do it yourself on this stuff. I'm actually—I'm biased. I'm a cloud guy. But I mean, there's some serious horsepower that goes behind training these models and all that stuff.

So, that could be a holdup as well when you roll your own, the cost of doing that and just trying to acquire all that infrastructure, get it set up, whereas if you were using the cloud services that's done for you and you can actually start focusing on your hypothesis there.

David Linthicum:
Yeah, it's going to be a big build-versus-buy decision that we have to make. One question I'd like to get your opinion on, I hear this from clients and from other people in the industry, that they're concerned that the focus on generative AI by the big hyperscalers is going to defocus them on the stuff that they run their business on. So, in other words, they're looking for innovation to occur on the storage and compute and databases.

Old school cloud technology has been around for 15 years. They're concerned about the shifting investment on generative AI, removing some of the innovation that's occurring on that side of the cloud, which is where their applications are living and will live probably for the next 10 years. Any merit to that?

Mike Kavis:
Obviously, it takes a different investment to invest in these things. It comes down to, do the things I'm going to produce out of this have more value than some of the traditional things I'm trying to maintain over here in the cloud? Yeah, it's going to change focus, but the question is, if I can get a lot of value out of this and then maybe let's focus on the value, not worry about how much I'm spending in the old way versus the new way. I've always looked at value.

If we're creating value out of it, great. If we're just playing games, not literally games, but if we're just playing around and playing around and spending a lot of money and business sees nothing, then yeah, let's focus on the bread and butter.

David Linthicum:
Yeah, it should be how we return value back to the business, always. And don't forget, you can ask questions and comment on this LinkedIn Live presentation. Mike and I will do best to respond, but we'd love to hear your opinions on this as well. And let's move on.

So, the other thing would be, I think, is going to happen, I'd love to get your opinion on this, is focus on cost optimization for cost controls and sustainability. And one of the things you heard besides the generative AI stuff in the cloud conferences last year was the O-word, optimization, your ability to get to a state where you're burning—you're able to accomplish the objectives of the applications and the data storage systems for the least amount of money.

And I think in 2022, as we talked about in the podcast before everybody got these big cloud bills and they said, "Hey, what the heck?" And so in many instances, there are three times what they thought they were going to be spending. So, I think we retooled for that with some of the FinOps stuff in 2023. I think 2024 is going to be focused on optimization. And really a byproduct of that, we also get to a better green ops scenario, in other words, we're consuming less cloud resources—on-premises resources as well, therefore, have a better carbon footprint and able to live up to the ESG capabilities and all these things are moving.

So, what do you think about that prediction? Do you think that optimization is going to be front and center and people trying to save some money? We already talked about why they would be doing it. They need to figure out how to free the resources up for the generative AI stuff.

Mike Kavis:
Yeah, I know just from our phones ringing that the amount of calls incoming exactly about this topic are like 10X what they were a couple of years ago and probably, 5X where they were last year. So, there's definitely a lot of attention there.

I think one of the reasons why cost is expensive is because the way we implement things in the cloud. And too many times, the cloud is just a data center to people. So, if you take what you have and move it to the cloud, well, the cloud is built for elasticity. If it's still running all the time—and in the data center, you paid for and it's on the books and then it depreciates over time. In the cloud, it's just like leaving your lights on all day long. And sure, it's going to be more expensive that way. So, it really comes how many podcasts we've been in, where I say, it really comes down to architecture. I mean, if you build things the right way, it can be extremely less expensive.

David Linthicum:
Yeah, I think so. And I think that optimization is really one of those problems that we really should have been solving all along. I know you and I have been screaming about stuff like that getting to an architectural threshold. We're able to build something that is going to be optimized, is able to, what we just mentioned earlier, return the maximum value to the business for the least amount of cash. And as a byproduct of that, you can build more sustainable systems as well.

So, I think it's just going to be a focus. I think it's going to be a little bit more of a momentum than we saw in 2023. I think 2023, everybody's moving to FinOps and retooling for systems, now it's going to be leveraging these things and then finding some ways to actually get to an optimized state.
The other thing about the—other thing we should discuss is the repatriation movement started in 2022 and continued in 2023 and where we think things are going to go in 2024. This is a bit overblown. I think it was more of a right sizing movement. In other words, people were realizing, well, this thing should never have been moved to the cloud in the beginning—to begin with. It's way under optimized for the cloud because we're losing our shirts on the resources, the amount of money we have to put into it. So, either we're going to change it, modernize it, get it to a good optimized architecture, either going to pay for that or we're going to put it back where we found it on-premises.

And so that's, I think, what drove a lot of the repatriation stuff. It was really fixing a mistake where they moved it into the cloud. Figured they would invest in modernizing it in the cloud, leveraging cloud native features, binding it to some of the performance advantages, ended up being this expensive beast because it wasn't optimized for the cloud. They weren't refactoring the system, they just did a lift and shift. And putting those things back on-premises was a big push in 2023. Do you think it's going to be same thing in 2024?

Mike Kavis:
Yeah, like you said, I think it's overblown. Any time a big company does it, it's a big article and it gets recopied 10,000 times. And the anti-cloud group will repost it five million times, but it happens. But again, most of the time, it happens for a couple of reasons. One, like you said, some workloads just never belong there, to begin with. The second, you just go in there without any of these FinOps controls or without any kind of housekeeping.

So, an example from my past, I was in a startup world and we got acquired, and the first thing they did was took away our access because it was a security concern. And then they took over all this stuff. And all of a sudden, the bill went through the roof. And one of the reasons why is because there was petabytes of backups that no one was—there was no process to go clean up after itself to no good hygiene.

So, quickly, the cost went through the roof. And the CFO came, said, I want to get out of the cloud. It's too expensive. And I go, well, time out, let's—why is it expensive? These reasons. Well, let's fix that. So, that's a lot of it. But some of it is certain workloads. Why even move it to the cloud? But most workloads can belong there.

David Linthicum:
Yeah. And again, it's going to be pragmatic look on the technology. And I mean, I hate to say I told you so, but we both told them so that these things had to undergo some work and some refactoring some changes to the architecture, and just shifting them into the cloud is not necessarily going to get you to where you need to go.

So, the other thing on 2024 is looking at the evolution around multi-cloud management and other complex systems that we're deploying, obviously, hybrid clouds, things like that. And continuing to focus on best practices and tooling to make that happen and getting to the supercloud, meta cloud stuff.
And it's funny, I posted a LinkedIn survey. You have these little quick LinkedIn surveys you can do. And I was like, "Well, what do you want to hear about?" And I wanted to get podcast topics. And I had generative AI, multi-cloud, and FinOps were the three choices. Pretty simple. Multi-cloud won out by a huge amount. It seems like, people haven't forgotten about multi-cloud in a light of generative AI. People are looking at that as their focus. That seems to be a problem that they're still looking to solve.

And I just see that as where people are going to invest their money in 2024 in terms of solving these issues. In other words, you got—there's always a downside to using multi-cloud computing. We talked about that before. You have heterogeneity, complexity, all these sorts of things should be mediated.

And so if you're moving to multi-cloud because you've taken a best of breed stance, which is fine, in other words, we're going to leverage the best services we can find on each cloud provider, therefore, we're going to end up with two or three cloud providers as the end state. The ability to manage those using common services, supercloud, meta cloud, common control planes around security and operations, and FinOps capabilities is something that we're thinking about in 2022, building a bit in terms of conceptually in 2023.

So, is 2024 where we're finally going to spend some money, and buy the AIOps tools, and buy the multi-cloud stuff, cross-cloud security systems to solve that issue?

Mike Kavis:
[LAUGHS] I don't know if any of these issues are solvable. What we can do is reduce the pain. But yeah, multi-cloud is a tough one because I think from most developers, their view of multi-cloud is, "I want to use the right tool for the right problem." And on the back end, it's usually, "How do I make this one experience?" And those two conflict, right? It's like, if I want this to be one experience, then you can't go deep in the clouds. You just have this one cloud that you come consume.

And so there's a lot of conflict there. And, really, we need to address both of these patterns. You need to be able to address patterns where someone wants to go all in the cloud, and you need to address patterns where someone needs to be cloud agnostic. And that's the battle there.

Too often, the discussions about multi-cloud are happening in a vacuum between these two groups. They're not working together. And that creates a lot of challenges. So, this is a really,—I mean, probably, the hardest part of all of this is trying to tackle the multi-cloud. All the cloud vendors are doing different things in different ways. At the logical level, they're creating the same services but different architectures, different APIs. It's very hard nut to crack right there.

David Linthicum:
Yeah, I think that you framed the problem well. I think what they're looking at is the ability to leverage this as a common framework and really, a planning kind of a problem. So, at the end of the day, all the—a lot of the other problems we're talking about here, it comes down to human beings and their ability to understand and their ability to plan. And I think that's where multi-cloud—if it's going off kilter, that's it.

So, in other words, people that have a complexity problem when dealing with multi-cloud, a heterogeneity problem, they can't operationalize it. They can't secure it because it's just grown out of their control. If you look at the root of that is they haven't done the planning that's needed to make sure they have the resources around to make it happen, the technology around to make it happen.

And when they talk to you, they normally just want tools. What's the best tool out there to run my multi-cloud? When, at the end of the day, if you put a couple of planning cycles in, you figure out how you're going to have a common control plane that goes across different cloud providers. If you figure out how you're going to—how you're going to leverage different technologies in different ways, are you going to mediate heterogeneity, all these sorts of things that are going to be part of it, that really is the problem to solve. So, people have to take a political stance to make this stuff work for them and do some planning, and then figure out the technology you need to bring in to back it into the solution. What do you think?

Mike Kavis:
Yeah. Yeah, the big challenge, though, is the more you try to control multi-cloud, the less you allow the developers to go deep in the stack. And that's where the real value is in this. So, I say, what things do we want to control? Operating system patching. Developers shouldn't be doing that, right? But if you want to—say, you can only launch from this UI, no matter what cloud you're in. And you can only do X, Y, Z. Now you're building the pass. And then my thing is, well, why don't you just buy a pass instead of building one?

So, it's tough. I really think you need to support both use cases. The use case where I can go all in one cloud and the use case where cloud needs to be abstracted and my workload needs to run everywhere. And that's so easy to say so hard to do.

David Linthicum:
It is hard to do. And I think we need to put some thinking behind it before we start tossing tools at the problem. And I think that it's what enterprises like to do. And before I see tool one show up, I need to see a plan in terms of where you are now, what your asset state, where you're moving to, business metrics, and all these—which it's hard to get people to do that. But that's ultimately the path is most likely going to succeed in getting some value out of this stuff.

Because I do see multi-cloud deployments pulling resources out of businesses because of the complexity, the heterogeneity, the amount of technology, diversity they have, that have to be mediated some way and you have to find common services to be able to run across these systems in order to get to a simple state. And that just takes a lot of work and planning. And it's fairly expensive. And I think that's why businesses are balking at it. But it's more expensive for you to move forward with overly complex solutions that are not going to be able to bring the value back to the business.

So, finally, the last one and also part of a what was a new concept, we're seeing what I've been calling ubiquitous computing. You wrote about this in Tech Trends at Deloitte, and the ability to spot the fact that we're not necessarily always in on cloud computing, moving to a cloud-centric kind of architecture, which I think we're looking to do for the last 15 years, to realization that we're going to deploy on multiple different platforms. It's going to be edge computing. Some stuff's going to exist in the data center, managed service providers, COLA providers.

And moving to a ubiquitous computing model, we're able to leverage the data and leverage the application where it exists and not necessarily forcing a migration to cloud, which I think was the big push. Certainly, the hyperscalers were all in on this. Everything has to be in cloud and cloud only, nothing but cloud, and all the cloud, and they have a never cloud people, never cloud, all that kind of stuff.

It's really the ability to accept that the destination of this is going to be a ubiquitous heterogeneous computing platform and the data resides lots of different systems and our ability to leverage the data where it exists and use it where it exists. For example, training data for generative AI systems is really going to be the path to success.

Do you agree with me? Do you think that's going to be a bit of a push in 2024 to figure out how to make that work and really moving away from having to force everything into the cloud?

Mike Kavis:
Yeah, I agree. And what's funny is I started writing about IoT must have been, I don't know, 10 years ago. And they weren't talking cloud. They were talking about processing on the edge.

David Linthicum:
I remember the Wi-Fi sheep you brought up in a meeting one time. Yeah.

Mike Kavis:
[LAUGHS] Yeah. And really, there's a need for both, right? So, the example I often use is the power windmills. They're processing a lot of data from sensors on the blades in real-time, and adjusting the blades to optimize throughput of energy. You don't need that in the cloud in real-time. But there's a lot of value in trickling the relevant data that's coming out of it, back to the data center to run all kinds of Machine Learning analytics to say, well, why does this wind stream or why does this type of weather impact the output of my machine? And then go make adjustments back out on the blade.

So, there's a good marriage of some data, making sense in the data center to do some deep processing, but there's always small data that you read it and you react to it. Same time with shipping things, logistics trucks, they're looking to make sure the temperature is right. All those things don't need to be in the cloud, unless you want to do analytics to figure out why certain elements create certain results.

David Linthicum:
Yeah. And also, what's interesting to me is the times I've sat on panels to discuss this have some emotional reactions to this. They just say like, if you don't say everything—that the cloud is the destination for everything, that's probably something they don't want to agree with. Obviously, people have different biases. But it just, to me, what we're doing is we're opening up our minds in terms of fact we're going to leverage many different technologies, including IoT and edge computing and all kinds of different platforms, which is always the destination. I don't know what this infatuation is with moving everything into a public cloud provider. Even the public cloud providers wouldn't agree that that's a healthy thing to do.

So, anyway, 2024 is shaping up to be a great year, Mike. So, where can the listeners find you on LinkedIn and other places? And how can they reach out to you?

Mike Kavis:
Yeah, it's just Mike Kavis, K-A-V as in Victor, I-S on LinkedIn. I don't do social that much anymore, but I am on Twitter, @madgreek65. I don't Tweet too much anymore. I've been overpowered with social media for too many years. But I'm there occasionally. But that's it. You can always look me up on LinkedIn. Reach out. Happy to reach back.

David Linthicum:
Yeah, make sure to reach out to Mike. You reach out to me too here on LinkedIn. By the way, ask comments and questions here. Happy to follow up with you if it goes past the time. And let me know what you think about the discussion. Let me know other things you want to discuss. Let me know what you think about the podcast, topics we should cover. We're interested in all that stuff, really getting to the point to helping people be more successful with cloud computing.

So, leave me a comment here. You can always go to our podcast website, that's deloitte.com/US/cloud-podcast. Check us out there. You can reach out to me directly through my email, dlinthicum, L-I-N-T-H-I-C-U-M at deloitte.com.

So, until next time, best of luck on your cloud deployments. We'll talk to you guys soon. Thank you very much for attending this LinkedIn Live event.

Mike Kavis:
Thanks.

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.