The Deloitte On Cloud Podcast

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

Title: **David Linthicum and Mike Kavis take a look back at the top cloud trends in 2023**

Description: In this episode, Deloitte's David Linthicum and Mike Kavis look back at cloud trends in 2023. Generative AI was the most momentous event in 2023, but there were other trends that played a significant role for organizations in cloud. Resilience, FinOps, talent management, and repatriation were top of mind for most organizations. As for the explosion of Generative AI, according to Mike, one thing is crystal clear: Cloud is absolutely the most critical component of a successful implementation.

Duration: **00:21:59**

David Linthicum:
Welcome back to the On Cloud podcast. I am joined by Mike Kavis, chief cloud architect at Deloitte Consulting, LLP, and today we're going to take a look back at the cloud and tech trends and wins of 2023. That's great copy. I wonder who wrote that. That's awesome. Hey, Mike, how you doing?

Mike Kavis:
I'm doing pretty good, Dave. Great to wrap up the year again like we've been doing for I don't know how many years.

David Linthicum:
I think we did a half-year wrap-up last time, so catch the listeners up as to what you've been up to and your role at Deloitte and what you've been working on lately.

Mike Kavis:
Yeah, I mean, a lot of what we're focusing on lately, my teams, are around resiliency. I think a lot of people, a lot of companies, have been in the cloud for a while, and they figured out at least good enough how to get stuff there and how to build stuff, and now it's, okay, how do we run stuff and how do we keep the SLAs up and stuff. So, a lot of focus on resiliency, whether that's in architecture, in incident response, or just pure ops. So, spending a lot of time in that space.

David Linthicum:
Yeah, it seems to be a big trend. You've got to think about it, we're probably 12, 15 years into the cloud right now, at least the enterprises are, and it's time to start thinking about stuff we need to keep this thing going long term, and in many instances, don't have a good resiliency program, maybe lacking some security issues, they don't have complete operational control of the systems, they're not managing heterogeneity as well as they can, so it's really—it really pays to double back and rethink a lot of these things and also

reconfigure them for better resiliency, better optimization, things like that. We're getting back into the basics of it. Is that your thought as well?

Mike Kavis:
Yeah, and some of that's the effect of the rush to get to the cloud, whether that was consolidating data centers or the COVID push. COVID made a lot of companies kind of highly prioritize getting there. So, a lot of it was just let's get there and figure it out later, and now we're at the figure it out later stage of that.

David Linthicum:
Yeah, we are. We are. I think it's double back and modernize and optimize, which is going to get us into the trends. So, we divided the trends up in two things, number one of course Gen AI related in the cloud—that's basically everything these days—and then talk about the non-Gen AI trends that were—that are out there as well, so lots of stuff occurred that wasn't directly related to Generative AI, so we'll talk about those trends which have been going on behind the scenes and many of what you just mentioned, the ability to think about resiliency and think about really kind of maintaining these things more longer term.

So, first thing kind of leads us to that. We're talking about optimization of the existing cloud-based system. I hear this a lot and we just had re:Invent and listened to the keynotes there, and of course the Google conference and the Microsoft conference talk about the ability to leverage cloud-based resources in much better, more optimized way, and I think it came to the issue that you just mentioned. In many instances, people sped too quickly to the cloud, certainly the pandemic drove a lot of that, and a lot of the systems are inefficient, even though they work. It's the whole it works stuff, that they may not work at the value points bringing value back to the business at the rate I think that cloud was initially sold at. What are your thoughts on that?

Mike Kavis:
Yeah, I agree with that, and some of it is really because a team moved some apps to the cloud that may have some dependencies back home in a data center that other people aren't aware of. So, I've seen a lot of outages where someone in a data center made an update to something having no idea that there was a dependency in a cloud app and broke it, and stuff like that. So, there's a lot of that too, just integration and complexity issues with legacy data centers connected to cloud-based applications.

David Linthicum:
So, do you think it's going to drive bad habits? Because if you think about it, we're fixing things retroactively. And, so, in other words, we missed something and under-optimized something, not making something as sustainable as it could be, as efficient as it could be, as resilient as it could be, and now we're going back and fixing these issues. Do you think the play is going to be under-optimization of these various systems, and because we're trying—the need for speed, getting the thing out there and, hey, we'll fix it later. Do you think that's going to generate some bad habits?

Mike Kavis:
Probably. I just remember my first job out of college we were migrating one mainframe to the other. We were on PDPs and old Burroughs systems, and we were migrating to this brand-new huge IBM 3090, and we didn't change a lick of code. We just took stuff that was built to use 8K worth of memory, breaking up the program to do that. We moved it to the cloud where we had this—back then we called it infinite memory. It's the same thing we're doing now, so it's just history repeating itself. A lot of companies a lot of pressure to deliver new stuff, and it's hard to prioritize technical debt. And I've been in this space almost four years and nothing's really changed on that front in a lot of cases

David Linthicum:
Yeah, and I think it should drive some change, but it may not And certainly with the next generation of Gen AI stuff, and we'll talk about that later in the podcast—it is going to pay to do the initial optimization of these various systems up front.

But I think enterprises are going to have to touch the stove to really kind of understand these things, and it is building lots of technical debt that has to go back. You have to go back and mediate at some point, and that becomes kind of a drag on value unto itself. So, if you're looking at getting the value of these cloud-based systems, it really should be a focus on doing it right the first time, but unfortunately, I think the business reality is that doesn't really lead to that and it's more popular just to lift it and shift it and get it running—get it out and running as quickly as you can.

So, the next trend I noticed this year would be FinOps, financial operations for cloud, the ability to kind of leverage these new systems—they're not new. They've been around since cloud has been around. So, in other words, we're looking at monitoring and accountability for these various systems and optimization of cost. But 2022 we got a big cloud bill, and we talked about this, I think, on the last podcast, where people were realizing that they weren't, in essence, getting the value out of cloud that they thought were coming. And you think about in the initial sell, a cloud computing CapEx versus OpEx is going to be cheaper, operational benefits, speed the business, agility, all this kind of stuff.

And then suddenly in 2022, certainly probably after the vast migration in the pandemic, we got these huge cloud bills, and then suddenly enterprise IT realized that they need to stand up certain processes and systems to allow us to have better accountability,

better optimization of these systems, be able to turn things off when they're not being used. I can't believe how much people are wasting on that, the ability to shift resources, things like reserved instances, get more value out of that, in essence, becoming a little bit better stewards on leveraging public cloud services which aren't free, they're very expensive, and many instances people are paying hundreds of thousands of dollars a month to get these services up. What are your thoughts on that?

Mike Kavis:
Well, I have some firsthand experience with that as well. Back in my startup days, we got acquired, and we went from a small team that kind of controlled everything to a large team, and a lot of people on that team didn't have cloud experience. And all of a sudden—because we did everything manually because we were a startup flying by the seat of our pants. Because there were no guardrails, there was no governance, all of a sudden, the cloud cost a bunch, and I had this CTO—or the CFO come to me and say we need to get off cloud, it's too expensive. I'm like, "Whoa, time out, we just need to put some guardrails in place because right now it's the Wild West." And that's the case in a lot of places if you don't—people are busy, people are pressured to get stuff out the door.

A lot of this stuff, the thinking should be done for them, the guardrails should be in place, the monitoring should flag things, you should be able to forecast, you should be able to proactively say, hey, my cost in this one area is going up, let's look at it just like you would look at an application that says, "Hey, I'm using more memory here. Let's fix it before it becomes a problem." So there's many services out there, there's so many apps, there's so much stuff going on it's almost impossible for humans to keep tabs on it without automation assistance and some guardrails on it. So, yeah, it's a big problem, especially like you said, people are 10, 15 years into it. There's so much stuff in the cloud now. It's kind of like when we—back in the day when we had an upgraded version of Windows, we had to do an inventory of what's in our enterprise, and there was stuff we didn't know. Well, now we have that in the cloud.

David Linthicum:
Yeah, it's going to be an ongoing battle and we need to start to gather the weapons now for making this happen in FinOps, and some of the FinOps tooling is starting to get a lot better. And I think just kind of linking this to the previous conversation, we're going to get back into how we can better optimize these systems and really kind of core metrics. And even if we have Gen AI systems—there I said it again—brought into the FinOps world, we're able to do some powerful things, like look at the way we're consuming and optimizing our cloud spending versus what our peers are doing and the ability to have training data from all these various businesses. So, we understand not only how we're doing, but how we exist in the marketplace and how we exist in terms of best practices because right now enterprises have no idea.

At the end of the day, there's not a mass understanding and not an automated understanding in terms of how to optimize this stuff, and I think it's going to be—it's kind of boring and exciting at the same time. But your ability I think to get more value back to the business from this stuff and really not doing so by sacrificing performance and capability is really, really opportunity here, and I think in 2024-2025 there's going to be a big push in that. I also link it to the next conversation if Gen AI's going to cost a ton and we're only going to be able to afford it if we're able to optimize the resources that these things run on.

So, next would be kind of shifting gears—talent management. And the reason I put that in there is I hear a lot of people talking about their success in cloud based on their ability to attract and retain cloud talent—cloud security architecture, cloud development, cloud FinOps—everything we just talked about, cloud resiliency. So, what are your thoughts on the trends here? Do you think that enterprises are going to become more successful at attracting and retaining talent or are they going to make some big mistakes?

Mike Kavis:
The companies that I've seen that have been the most successful are the ones that have built internal "colleges" or training programs for cloud, and they supplement it with all the third-party solutions out there too, but they make investments in training. So, I think the companies that are doing well here really know that, especially with Gen AI coming, everyone needs to be savvy, whether you're an engineer or you're a salesperson. So, the companies that see that will succeed a little better.

David Linthicum:
What about leveraging training technology in different ways? Are we moving into an era where the on-demand training is going to morph into something that's even more sophisticated in the ability to train people up very quickly no matter who they are, where they are, what they're doing?

Mike Kavis:
Well, I think AI is changing training. Training is a little less instructor-led now and it's more—at least for the technical folks like us, it's more interacting with an AI bot. So, one example I saw the other day is like myself, I've never built a machine learning app, but I could go into Code Whisperer or Code Pilot or one of these and say what are all the things I need to do, and it gives me a list. And I could interact and finally get to a place where I'm learning on the fly but getting code snippets and getting direction on how to build it. So, as opposed to attending a class for three days and then trying to remember it all and then go apply it whenever you get an opportunity, this is real time interactive learning on the fly, and I think that's—even the education programs are heading that way as well, so I think the way we learn is going to drastically change.

David Linthicum:

Yeah, absolutely. And I think it's going to be on the enterprises to figure out how to do this in an effective way. Finally, return to on premises. We talked about that last year as well. Let's kind of get a read on where we are, and I kind of don't look at it as—and even people say mass amounts of repatriation because the cloud didn't work. That's not the case. I think what happened was, in many instances, these applications and data sets shouldn't have been moved to the cloud in the first place, and if they're not willing to make the modernization investment to leverage more cloud native features, then pushing it back on premise may be a better, cheaper option, at least in the short run, but what's your take on the whole repatriation thing?

Mike Kavis:
Yeah, I'm a little biased on this because I started my cloud journey building all in on a single provider with five guys and competed against large companies and blew them away, so I'm very much an all-in cloud person. When I see people bringing it back, I don't think it's a problem with cloud. I think it's a problem of unsuccessfully implementing whatever they were trying to implement, whether it's, as you say, that app shouldn't have been there or even if it should have been, the way they implemented it was wrong. So, going back to my earlier example, costs went through the roof because there were no controls in place and the CFO wanted to get out of cloud. I'm like it's not a cloud problem; it's an execution problem. And that's a lot of it.

David Linthicum:
Yeah, it's a huge amount, and I think most of these things are self-inflicted, and it's really not a push against cloud or even a push against on premise. Those things are both going to be an option. You have to make the investment in refactoring, modernizing these applications if you're going to use it in the cloud in a more efficient way. So, let's talk about Gen AI in the cloud. So, this has been kind of a re-revolution in the middle of the cloud evolution, so to speak. AI's been around for a long period of time, and you have to be under a rock not to see the whole Generative AI revolution that occurred and really kind of people seeing the amazing potential for this technology to transform businesses and even create businesses moving forward.

But there's always some things and concerns we have to think about, and so even with the explosion of Gen AI when you do the costs and even the back-of-the-napkin calculations, this stuff's going to cost a lot to run. They're very process intensive, data intensive, they take in many instances specialized processors which they're going to charge you more for. So, the focus was on cost of Generative AI. Do you think next year we're going to be able to solve some of these concerns?

Mike Kavis:
Well, it depends on the approach companies take. A lot of companies want to be cloud agnostic, so they're going to try to tackle this themselves, and the amount of infrastructure it takes is incredible. If they buy into the cloud provider's approach, it's pay as you drink type thing. you can scale that with your costs, but most companies these days want, for security and privacy reasons, want to do this themselves. And after seeing what it takes just to stand up the infrastructure and train a model, forget about what you're trying to deliver to the business, just to get that in place so you can start building that, the cost is insane if you do that yourself.

So, to me, I think if the companies that leverage the cloud providers will be more cost-effective than the ones that are going to try to do it themselves, and every—a lot of companies are trying to do more with less, so I don't know how you can go to the board and say I need 5,000 nodes in a cluster to do a proof of concept. That's hard, so if there's ever a time to go all in on a cloud, this is it. The amount of horsepower required to maintain and to patch and all this stuff is crazy. Probably more than they already have to do the rest of their business.

David Linthicum:
Yeah, and the reality is it's going to be very expensive in the cloud and very expensive on premise. It depends on which one's going to be least very expensive. And I think as—to your point, not only do you get some cost advantages in moving to the cloud, but also you get some strategic advantages in that they're building and maintaining the AI infrastructure there and they're providing the core integration systems, so you're not having to DIY everything, which is where I think things go off the rail when people start building things themselves, certainly now on modern systems. So, you don't have to maintain core updates and patch fixes to the latest Generative AI systems you're moving out there, your ability to leverage different kind of data models, and you're able to take training data offline and put it on slower base systems and really kind of provide the optimization to make it happen, but the expense is still going to be there. It's still not going to be cheap.

But I think if you're—again, back to the previous conversation, you're making good optimization, good decisions about what platforms you're able to use and looking forward not only one year but five years because you're going to have to maintain it long term and build it right the first time and build it efficient the first time, I think that's going to pay back dividends and we're able to do things faster and hopefully cheaper and just becoming more efficient in how we're building and deploying these architectures. So, you think we're going to get smarter about this?

Mike Kavis:
Probably not. History repeats itself. And you can't just take any old person to go stand up huge petaflop-based clusters. You need specialized skills, so you've really got to leverage the cloud for this, in my opinion.

David Linthicum:

Yeah, absolutely. So, shift our focus then to 2024, looking ahead a bit. What are the likely Gen AI trends in 2024? Gen AI trends in the cloud that you'll see as really kind of moving the market and getting the attention, the press, and the enterprises out there who are consuming this technology?

Mike Kavis:
Well, there's a bunch. A lot of it is pushing easier decision-making to nontechnical people, so the IT professionals create the infrastructure, the processing, but it puts data at the fingertips of marketing people, salespeople, businesspeople who can interact with whether it's chat bots or with push-button dashboard creation. We saw a nice demo of that the other day where one person, not a team, quickly whipped up an app and at the end of the app, they hit a dashboard button and it went through all the data and created executive dashboards based on the execution of the program and the ROI and all that stuff.

So, I think that's one angle where we're putting more data and more intelligence - artificial intelligence in the hands of nontechnical people so they can do their job without waiting six months. And then the other one is how we actually build code. Working with the chat bots and some of this stuff will actually help us refactor code, help us make architectural decisions, help us debug, help us test, so accelerating the whole software development life cycle and making it easier to do our job.

David Linthicum:
Yeah, I think so, and I think you said something very profound last time we did the podcast together was that we're going to be focused on people who are able to drive the prompts, the ability to ask the right questions to these systems is really going to be a core skill moving forward. It's not how we build all these things in the background. I mean, generating applications and coding stuff, those are all tactical things that can be done by certainly skilled humans and now we can do it by AI, but your ability to ask the correct questions and your ability to understand when you're getting the correct answers back from the system is really going to be a core skill in 2024-2025. Do you agree?

Mike Kavis:
I agree, and it's kind of Sherlock Holmes type thing. You're actually asking questions, drilling in, you're solving puzzles to get the chat bots to produce what you want it to produce. I can go say give me code to build an app, and it's going to give me some code, but I can also reference it with this is the style of programming that I like, this is—I want you to do test harness. You start describing exactly what you want, and you start getting better results. Now we're getting tools where you can provide that context of your business, of your company, and it can start helping you interact and solve those problems there. But if you just ask blanket questions, you're going to get blanket answers. If you can figure out how to iterate with that chat bot and get to better questions, those are the people who are going to get a lot of value out of it.

David Linthicum:
It certainly is a skill that a lot of people need to have. So, where can we find your stuff on the web and also your books?

Mike Kavis:
Yeah, so I have a couple books out. One of them is actually pretty old now. It still applies, *Architecting the Cloud*, I think back in 2012 or '13. It's way back, and you reviewed that one a long time ago, couple companies ago, I think. And the latest one is more on operating models. All that on Amazon. On Twitter, I'm @MadGreek65 and Mike Kavis on LinkedIn.

David Linthicum:
Yeah, make sure to follow Mike. 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, who's on the blower right now. Find out more at deloittecloudpodcast.com. If you'd like to contact me directly, you can email me at dlinthicum@deloitte.com. So, until next time, best of luck on your cloud journey. Everybody stay safe. Cheers. Bye.

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.

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