The Deloitte On Cloud Podcast

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

Title: **The CTO Advisor's Keith Townsend on how to make good decisions about Generative AI**

Description: Generative AI couldn't be hotter. Everyone wants in on the revolution. In this episode, The CTO Advisor's Keith Townsend agrees that Generative AI is revolutionary, and he discusses cloud's role in Generative AI, the decision-making process, and potential pitfalls around implementation. However, he also urges companies to base their Generative AI decisions on a trusted maxim: The value of a given project is proportional to the value it returns to the business.

Duration: **00:26:01**

David Linthicum:
Hey, welcome back to the On Cloud podcast. Today on the show I am joined again by Keith Townsend, Resident Global Technology Advisor at The CTO Advisor, a Futurum Group company. So, that's a bit of a change, Keith. Update us with that.

Keith Townsend:
Yeah, so back in October, I had been talking to Daniel Newman, who's the CEO of Futurum Group for probably a year and a half, two years about doing something special. And back in October 2023, we decided to go ahead and merge companies, and now we're part of Futurum Group. A couple of weeks on the ground, we soon realized that we needed to kind of tear up what we originally thought because this combination's been much better than what we thought it would be. So, now I'm basically head of advisory because of my experience in management consulting. It's kind of peanut butter and jelly.

David Linthicum:
That's awesome. I always like following Daniel on LinkedIn. He probably posts five or six times a day. Sometimes it feels like five or six times an hour. He's at every conference.

Keith Townsend:
I'm here and I honestly still don't know how he does it. It is amazing.

David Linthicum:
You've been doing some really awesome videos. I can always tell they're awesome because I always watch them all the way through, and those are hosted on LinkedIn, is that correct?

Keith Townsend:
Primarily LinkedIn if you're not X.com. I do it on X, but primarily LinkedIn is probably the best platform for engaging in our area of focus, right?

David Linthicum:
Yeah, I think so. I think so. Videos, I guess you put it as many places as you can, put it on X, put it on YouTube, put it on LinkedIn, and people seem to find them wherever we want to find them, but it's interesting. But you go into details, which I like, and you go into kind of have a very inside baseball conversation on how to use this technology and mistakes people are making out there, how they can do things better. It's kind of the heart of a teacher stuff that I think gets me. So I urge everybody who's listening to this podcast make sure to go follow Keith on LinkedIn. And normally we say this at the end of the podcast but do so now. Follow him on X and watch his videos. Even watch the ones he's made months ago as well as weeks ago because I think they're incredibly valuable.

Keith Townsend:
High praise, David. I appreciate it.

David Linthicum:
Yeah, it's worthy. So, anything else in your life you've been following? What are you focusing on these days? Obviously, Generative AI is one thing, but anything else? What's been your area of focus?

Keith Townsend:
So, our primary area of focus is still hybrid cloud, so I'm really interested in where Generative AI and public cloud begins and where the private cloud starts to make sense. I know you've written some great stuff on this in the past couple of months. It's a high area of interest, but our primary focus is still the hybrid cloud.

David Linthicum:
Yeah, and I think it's going to be that for a long time. If you look at it, I think the writing's on the wall. This is just my opinion, but we're definitely moving to a heterogeneous, ubiquitous state where I think a few years ago everybody was saying, "Hey, we're going to consolidate on cloud, everybody's going to move to a major hyperscaler over the next few years," and now we're realizing that that's not always financially the right decision, and also from an architectural point of view, that might not be the optimized platform that we should be leveraging, so things are getting complex. Cloud's not always going to be the answer. Hardware has come down a lot in price, almost a 45-degree angle in the last ten years, and so enterprises are out there with this Generative AI opportunity, which means we're able to create game-changing systems that are going to define the next generation of the business, and that's truly what's at stake here.

Companies are going to be able to create systems using this technology which defines who they are, very much like the rideshare programs and the house share programs, all these sorts of things where they don't own an asset, but they own the IT. But the ability to do this in pharmaceuticals, the ability to do this in manufacturing, and the ability to kind of bring better customer experiences and more efficiency and optimization to a supply chain leveraging technology and leveraging data that really is now accessible now thanks to the commoditization of it and also the rise of Generative AI infrastructure that's out there now. Give me your take on it.

Keith Townsend:
Yeah, so a great example of that is the availability of GPUs versus CPUs, et cetera. These used to be very black-and-white decisions. If you do the math of GPUs on-premises versus in the public cloud, the math is very compelling that you just buy GPUs if you can handle the power and cooling requirement. But the availability of these limited resources highlights the dance between the public cloud and private infrastructure, the ability to go to a centralized resource, consume those centralized resources as needed, and then when scale and logistics accommodate it to bring that stuff back on-prem. It is the lifecycle probably accelerated that we've seen over the years. When you want to move fast, I know you've written and talked about this—you do it in the public cloud. Then when you want steady state operations, controlling your costs, controlling your flow of data, you bring that stuff in-house. Generative AI seems that on steroids.

David Linthicum:
Yeah, and I think it's going to be the "it-depends" decision that we're making on all this stuff, and I think it's—and by the way, GPUs aren't always a slam dunk for Generative AI systems. Most of the Generative AI stuff we're going to do is not going to be ChatGPT size LLMs. It's going to be very small—people calling it small language models, whatever you want to call it, it's going to be a tactical use of the technology where it is automating inventory control, it is automating supply chain replenishment, all these sorts of things which are really kind of more important to the business than having to write thank you notes and write a song about yourself.

And if you look at the underlying infrastructure, GPUs aren't always going to be needed, definitely not in an edge scenario. In other words, we're running Generative AI systems—and I was working on an architecture the other day for this. Out on an edge platform where you're going to have 10,000 of these, things that are installing cars and equipment in aircraft, things like that where it just doesn't make sense to have GPUs hosted there, and our ability to kind of use alternative processors or commoditized processors really is something we should be looking at. What are your thoughts on that?

Keith Townsend:
So, I think a great example of this, and something to do a deep dive on, and we'll do this at one of our events, I talked to a greenhouse farming company up in Canada, and they're doing some amazing stuff around sustainability. They're burning clippings from their produce to produce energy, piping that carbon back into the greenhouse, really great sustainability approach to farming. But they've always used data to enhance their crop yield.

For example, they put an IOT sensor on every tomato plant. And we're talking about hundreds of acres of farming. And the scale of that data set is just mindboggling in itself, but they're also using this data to determine when is the best time to pollinate a flower so they get the perfect bell pepper. This is all AI/ML work happening on CPUs. And as you talk through kind of that detail inside-baseball conversation, I asked them why not GPUs. And the VP of IT shared with me how GPUs break their firmware upgrade, their software lifecycle, their governance that they've optimized for this really lean operating environment.

Sure, GPUs would probably save them four hours, five hours of time every time they run a job, but they only run a job occasionally, and CPUs fit their operating model overall with their VMware, V-spirit environment, being able to V-motion workloads and balance their overall clusters. They can't do that with GPUs. So, these interesting governance and operational challenges when you get on the ground become really, really clear that this is not a black-and-white, "Hey, we use GPUs for all AI and our general compute," we kind of just forget about that, which doesn't happen in the real world.

David Linthicum:

Yeah, and if you look at the cloud conferences we've had in the last year, which basically were Generative AI conferences. Obviously partnering with GPU providers and all those sorts of things are exciting and there's a massive run on GPUs and so there's shortages of them, and of course the prices go up, and now there's other innovations in the marketplace to create one-off processors that are bespoke for Generative AI which are able to leverage some patterns of GPUs but burn less power and cost less money. So, the thing is, if you look at the pragmatic use of this technology today, and that's what people forget. In other words, we can't spend $1 million a server, not that we ever would do that.

Or we can't have a cloud bill that's $200,000 a month to support these systems if we're going to have value that comes back. So, using architectural principles that we're going to leverage a platform that's going to be optimized for that particular purpose, in many instances, we need to look at the commodity storage systems, and we need to look, in some instances, at on-premise storage systems and edge-based computing, things like that. So, we have to look everywhere for where there's technology. So, you've got to remember this is going to be monster applications, even in their small form. They take a tremendous amount of data.

The entrance engines eat a lot of processing power, huge amounts of storage you're going to need for the training data, we have to have a huge amount of network bandwidth to bring training data in from the outside. All these resources are going to be leveraged. Now, here we are maybe 12 years, 13 years into cloud computing, and if people made mistakes with cloud computing, they didn't optimize the platform and the utilization of the platforms in the way they should. They either overprovisioned. They didn't refactor their applications which were moved to the cloud, which was the big one. So, they moved inefficient applications into the cloud, expected some magic would occur, it didn't occur. And now we're going to have an opportunity which I think is the most exciting. We're going to build net new applications that are truly going to be game changing.

And obviously the instinct there of the architects is let's get the best, fastest platform, which is also going to be the most expensive. If they do that, this is going to fail. We're going to end up spending way too much on this infrastructure, way too much on the development tasks of it, and I think at the end of the day, we're going to end up holding lots of assets, some we actually own and some that exist in the cloud where applications are bound to that asset. They're going to be way too expensive, and therefore under-optimized. And people aren't thinking like that now. What's your take on this?

Keith Townsend:
So, David, we've both done this level infrastructure, infrastructure planning for, between the two of us, probably well over 50 years, and what's old is new again. These challenges are not new. I don't think we've experienced the challenge quite on the scale of the Generative AI hype, but I think both of us can point to a decade or so ago when SAP/Hana, and other in-memory databases were hitting the enterprise. There was kind of this unknown return of value in the speed of innovation happening and in-memory databases were outperforming what obviously cloud providers at the time could do. And even it broke our internal depreciation schedules. You could not depreciate a system for five years, let alone seven for this type of application. You were replacing these systems every two and three years.

So, we had to figure out from both a finance perspective, from a CapEx versus OpEx within finance, from a data center refresh perspective, from a licensing and negotiations perspective, how to handle this shift in computing. We are now in probably another generational shift. I agree with you that I don't think customers absolutely know what the return on value of their AI projects, especially Generative AI projects will be. Will having a chat bot really move the needle for me financially? And I think you wrote a really great piece on how the performance aspects of Generative AI. I think a lot of people are taking back by the latency of chat bots, et cetera.

As technologists, we can appreciate how much processing goes into inference and the ability to get these intelligent responses, but from an end user perspective, we're accustomed to search engine types of response times, and turning that knob to get the right infrastructure to match the user experience and the business value has been extremely frustrating to find that balance. So, yes, we're in a new era of figuring out sizing our environments for the return on value, I think the important thing is for customers to do is to just sit back, slow down a little bit, and ask these fundamental questions that we've answered in the past.

David Linthicum:
Yeah, I think you just hit the nail on the head. I think we need to pump the brakes on some of this stuff because we're going headlong into making many of the same mistakes we made in the early days of cloud computing, we made in the mass migrations during the pandemic, and not having the architectural forethought that we need. And the problem is that we're going to get to a state where we're binding our particular application to a Generative AI engine which is bound to a processor which is bound to a cloud. And by then it's too late.

The relocation costs, you're locked into that platform, that cloud platform, because you're leveraging whatever native APIs they're using, whatever databases they're using. You can always pick it up and move it someplace else, but it's going to be extremely expensive and economically unviable to do so. So, right now in front of us in 2024, as we're doing the planning, I think with the state of the industry as most people haven't implemented large-scale Generative AI systems, they've certainly played with it, and everybody's hit the Generative AI systems out on the web. And certainly are looking for the applications. And now suddenly this new AI—this new executive AI officer is a new title that I hear everywhere. It's just like executive cloud officer a while ago. And they're making core decisions that are extremely important to the point that I think they're going to, if they're improper decisions, kill the business. They're not going to have the innovation they need to grow into their market when their competitors are able to do or they're going to overpay, overspend, overutilize this technology.

And I think pump the brakes is exactly right. In other words let's take the planning cycles and look at the future and get lots of different opinions to make sure we're making the right calls here because a lot of these things are going to be hugely expensive if we don't do that. And the reality is everything works. The end state of all this is I can build a Generative AI system that works, but I can either do it for $10 million or I can do it for $1 million, and I want to do it for $1 million and bring more value back to the business. And if I do it for $10 million, guess what, it still works. That's why there's no metrics in failure because all this stuff is going to get to a successful deployment state, most of it, and I just don't think that either there's lacking talent out there, lack of experience, lack of education in how people need to properly use these platforms. So, what would you consider the best practices? So, if I'm at a conference and I come up to you, and you always get this question, hey Keith, where should my Generative AI system run, how would you answer that?

Keith Townsend:
I would say start where you're at. If you're not clear on the business value and the return of a Generative AI system—and a lot of people are getting this frustrating directive from board level. We need a Generative AI strategy, but they haven't defined what the true application or value prop is, so this gets to kind of the technologists' AI expert level. And they're asking the question, well, should I size something with 12,000 GPUs? I was at OCP recently, so to reinforce one of your earlier points, Meta got up on stage and talked through how they're redesigning their network to take advantage of all compute and not rely heavily on GPUs. I think they just announced they bought 350,000 GPUs, but they still reworked their network so that they aren't completely reliant on GPUs.

And I think that's where I would advise customers to start. If you have an unknown set of requirements, I would get your governance and workflow around AI, Generative AI set. If you have some excess—X86 compute, start there. There are models, there's workflows that take you from X86 to GPUs. All the cloud providers support that workflow, so when you do need to expand, you won't be caught behind the eight ball and not knowing your processes and how to answer the internal question. But develop the expertise, and expertise can be developed on your existing systems.

David Linthicum:
I love that answer. Start with what you have because I do hear people, and they've reached out to me directly, we just bought $2 million worth of GPU powered services that we're going to install in our data center, co-lo provider, what have you, managed service provider in some instances, and they don't know what they're building yet. And that seems to be the reaction. It's like cloud was the same way, but cloud's fixable. In other words, if you overprovision the cloud, you can always bring it back, but these things you can't sell them. They're your equipment, they're capitalized, they're going to be depreciating, they need to be put into a value place.

And I think that ultimately reckless decisions are going to end up being an issue, and I think that the best practice is to start with what you have. This is not cloud versus on-premise versus edge. This is you understanding your own requirements. Figure out the systems that you're going to build and get value from. And by the way, that's not going to be everything. If someone says it's everything, they're not giving you the right answer. It's going to be certain uses within your business, for example, pharmaceutical company, your ability to deal with an allergy dependence system, so you have an expert system that's able to spot that, which provides a better customer experience, great service. That's an example of it because it's able to use not only our data, but the petabytes, zettabytes of information to figure out patterns of use.

And it's not going to be for inventory transactional systems. It's not going to be for sales transactional systems. If you're using it for that, then it's probably the wrong use. It's going to be just a few applications, I think, that are going to be the killer use cases within most businesses for the utilization of this technology. So, I think that the two things we have in front of us now, the challenges would be overutilization, in other words, I'm sorry, not having the correct use cases for this. You brought that up. And also overestimating the amount of money that needs to be spent and not moving the architectures to the minimum viable set of systems that are going to be optimized for the business.

So, we keep trying to think differently in how we do this, and I think it is going to be a planning cycle where we win, and also it's going to make things better for you because the innovation's occurring like gangbusters right now, and so you'll be able to see the patterns of where that's going. So, what would your final advice be for someone who's going to do this planning cycle? How do we staff up for that? Who do we need to get advice for? What kind of talent skills do we have to have within the organization to be successful?

Keith Townsend:
So, I don't think I can overemphasize the point that both you and I are making. It is a steep learning curve to move into AI. It is very bespoke right now, and you think kind of early days of Kubernetes and getting Kubernetes systems up and how simple—how relatively simple they are now. These systems are extremely fragile. You need the right dependencies, et cetera, et cetera. Develop that skill. That muscle, developing that muscle will literally save you millions of dollars down the road. Not every organization will do training models.

You don't need 30,000 GPUs in order to create a model from scratch. You're going to pull a model, whether that's GTP4 or some other open-source model, you're going to pull that down from the web, fine tune that, do some rack training on top of that. If you don't know what any of those terms mean, start there. Before you spend a dollar on AI, learn what an ALLM is, know what RAG is, know the difference between fine tuning a large language model and training a large language model. That would then put your mind in the right mindset of sizing your overall infrastructure requirements.

David Linthicum:
Absolutely. Great advice. I think there are specific things you need to understand about AI engineering, and you've got to remember these are complex distributed systems with many different components. We have the inference engines, we have the model development, model tuning, but we also have the database aspect of it, the output of the database, input in terms of training data which is always going to be different. Those are going to be widely distributed. So, we're going to build these systems using patterns I think we haven't used yet.

We've had AI engineering for a long period of time. I did it when I was 18 years old, but right now, in order for these systems to do their job, they have to have many different moving parts. It's highly heterogeneous. You're going to use different technologies, different platforms, and your ability to kind of ask yourself what those things are and your ability to do the correct things in engineering the platform is going to successful stuff. So, learn all you can right now. I can't stress that enough. Let's pump the brakes if you don't think you have the skill sets around to make it happen, focus on planning, splurge on planning, get the education you need, and use whatever resources you need to get things to the next level. And you've got to remember there's huge amount of resources in the businesses on the line with this one. So, where can we find more about you on the web, Keith?

Keith Townsend:
You can find me on the web still, the blog and the content engine still exists, thectoadvisor.com, on X I'm @ctoadvisor. And obviously, as we've talked before, I post an awful lot on LinkedIn.com.

David Linthicum:
Yeah, I urge you to follow Keith wherever he posts because it's absolutely a wealth of knowledge that he gives away for free and understanding how to use this stuff. And what I love about it, it's always insider views of it. In other words, it's not what the marketing people are saying. It's what this stuff really means, how you can use it, pragmatic advice, and I think we need more of that out there.

If you enjoyed this podcast, make sure to like us, rate us and subscribe. You can also check out our past episodes, at deloitte.com/us/cloud-podcast. Now to share a personal note with all of you, although you will be hearing my voice in upcoming episodes of the On Cloud podcast, I am officially stepping away from the microphone. The show will continue to bring the inside stories and unique perspectives you've all come to expect from the On Cloud. Thank you for being such an indicial part of this journey, I encourage you to keep tuning in. Best of luck with your cloud projects and stay safe.

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. Copyright © 2024 Deloitte Development LLC. All rights reserved.