The Deloitte On Cloud Podcast

Mike Kavis, Managing Director, Chief Cloud Architect, Deloitte Consulting LLP

Title: **AI unleashed: Deloitte's Mike Kavis explores the power and impact of Generative AI**

Description: In this Knowledge Short, Mike Kavis offers a primer on Generative AI. He gives a concise definition of Generative AI and discusses possible use cases and their benefits. Next, he dives into how Generative AI does its job, discussing how AI models learn and create content, as well as the processing power they need to function. He also explores the impact of Generative AI on business performance. Finally, he touches on hot-button issues like data security and AI and human collaboration.

Duration: **00:18:31**

Mike Kavis:
Welcome to the Deloitte On Cloud podcast, where we get real about cloud technology. Today we're going to do something a little different, what we call a Knowledge Short where it's just myself talking about hot topics in this space based on being a practitioner and an analyst out in the field and what I've learned and what I see. So, I'm your host Mike Kavis, Chief Cloud Architect over at Deloitte, and today's hot topic, nothing else but Generative AI.

So, let's start with some numbers. I was in a meeting the other day. Our global head of Generative AI and his team were sharing some numbers that they're seeing. And let's just start with where this is going. So, here we are, 2024, very, very early in this space, and predictions are by 2026 80 percent of enterprises will use Generative AI in some form or fashion. And just so you – just to expand on that, whether you want to use Generative AI or not, all the vendors that we work with, whether it's our e-mail office products, whether it's our CRM solutions, development tools, they are embedding, embracing Generative AI. So, even if yourself or your company doesn't actively go out and pursue Generative AI, the tools that you use are bringing it to you anyway. So, I think that's one of the reasons why this number is so high, is you're just going to have Generative AI at your fingertips at all times as we progress into the future.

Another interesting number: 2027, 16 percent of IT budgets will be dedicated to Generative AI type resources/projects, which is a pretty significant number. But wait till you hear some of the bigger numbers coming out. But first, 2028, the prediction is that 20 percent of all automated processes out there will be running under Generative AI. Generative AI will be used to automate all these processes.

And then, here's a mindboggling number. By 2030 – that's only six years away, five and a half years away – 100 percent of IT budget will be directed either directly or indirectly at Generative AI. And again, the indirectly part is you're consuming services that inherently use Generative AI inside. So, that's a bold prediction, but basically by the turn of the decade Generative AI is going to be everywhere. So, keep that in mind.

So, before we get into some of the topics that I want to cover near and dear to my heart, just want to get some basic terms out there and make sure everyone's on the same page. I apologize if you're advanced and this may slow you down. But let's get everyone on the same page of what all these terms mean.

So, what is Generative AI? There's a million definitions out there. I try and keep it simple. But it's an artificial intelligence that you use to create a variety of content. We'll keep it simple like that. So, a learning engine that creates content. And what kind of content does it create? It can create text. It will help you write things, whether it's an article or marketing material. It will help you write code. It could be video. It could be audio. It could be images. I myself am a large user of Midjourney, which is an image generator. I create, but I have no artistic ability whatsoever. I can't even draw a stick man. But I have a very

creative brain and I've created some amazing, amazing images that there's no way in the world I could create pen and paper. So, that's another cool solution.

So, it's just basically creating content and using a lot of technology, which we're going to talk about in the next few terms. But what it's really good at is it's learning from the data. So, it creates all this content but it's doing it based on learning from existing data, and then it uses that data to create new content. And it can create some very high, complex content such as our predictive models. It can help you build software products. There's really no limit to what we can do with Generative AI. And it's really important to note that we are in the infancy stage of this. The baby is still in diapers and hasn't even started walking yet. What are we going to be able to do in 2030 when it's 100 percent? So, someone said in the conference I was in the other day, I mean, Generative AI is very early. It's got a long ways to go. And it's already amazing, the types of things we see coming out of this solution.

So, a couple big terms that get thrown around a lot: LLMs and GPUs. So, I'm not going to go into great detail but LLM:, large language models. And this is really the engine behind Generative AI that processes all this data, learns from it, creates the insights and all those types of things. And I think what's important there – ChatGPT was the first big name Generative AI solution we started hearing about a couple years ago, and it's using an LLM under the covers.

Going forward, just like when we first embraced cloud, the public cloud, there was a lot of apprehension about certain corporate data being out in the public. You're seeing the same thing here with LLMs. A lot of companies are like, "Well, for certain information within our walls I want a large language model in our walls so only we can consume it." And that's what's happening now. Right now, there's a ton of work going on: How do we create these LLMs? And there's industry-specific ones; there's product-specific ones. Using some of our office products, under the covers eventually we could be loading in our organizational policies and controls and security and all that stuff in there so as we access these models it's within the constraints of our organization. So, there's a lot going on there. There's a marketplace for these types of things. So, certain products or even services have custom LLMs to address that particular problem they're trying to solve. So, that's cool

And then, GPUs: graphical processing unit. And what's amazing about this, some of the big vendors are producing these GPUs, which can perform way faster with better efficiency than CPUs. Nothing wrong with a CPU, but CPUs are typically built with 18 to 16 cores do basic computational needs that we've been processing all our lives. And GPUs are more tailored towards working on large tasks in parallel and will have thousands and thousands of cores and it distributes the work, which is much needed when we start to think about large language models, artificial intelligence, deep learning, machine learning. It takes massive scale to do these types of things, and your traditional CPUs, it would just take so many more processors to do what a GPU can do, because GPUs are specifically tailored for solving these types of problems.

So, those are some of the basic terms that are thrown around a lot when we talk about this space. I just wanted to get it out there.

One thing I wanted to talk about, though, I had a conversation with a few colleagues here. Being in IT and growing up in development and crossing the bridge over to the Ops when I got in the cloud space, there's a lot of great solutions out there for monitoring, logging. And one could argue that all these problems of predictive analytics and proactive learning have been in these tools for a very long time. I personally was first exposed to some of these tools around 2010 or '11-ish in my startup days. And these are amazing tools, and we were able to do a lot of things proactively.

I guess what the difference is from my conversations with my peers and colleagues is what we call supervised versus unsupervised learning. So, back in the day I would go into those tools and set up what I call units of work. So, let’s say my business is a travel website, and a unit of work might be the process of searching for a flight, finding a flight, booking a flight. And there's a bunch of services in there, and collectively I can put that in as a unit of work and then tell the monitoring system, "Go monitor the average time of all these services collectively in this unit of work. And then, send me an alert when it's X percent off."

So, that's an example of supervised learning. I already know what I'm looking for. And these tools are great. I give it that information. It can track. It can trace. It can get averages. And it can basically serve up my need. But things are still breaking on me. And why are things still breaking on me? Because I haven't discovered the next problem yet. And that's where these tools kind of come in. These tools exist already, but that's where the AI in these tools come in, is now it's talking about the unknowns. And it has all the data from our systems and it's starting to learn patterns of things that we haven't known yet. So, it's even more to the predictive side than before, where before we were trying to be very proactive; now we can be really proactive because the system can alert us of our unknowns. So, that's a big difference.

One of the first questions I asked two years ago when I got in this, and I'm sure a lot of people are asking is, "I already have all this. These tools already do this." Well, they do it in a supervised fashion but now we're moving towards unsupervised. And when you start talking about customizing, they're all embedding LLMs in their solutions, and soon we'll have the ability to customize those, and it's just going to get better and better. And like I said early on, the technology is very, very early. The vendors are very, very early in embedding this into their solutions. It's just going to get better and more efficient over time.

So, one of the other topics, I hit on it a little, but it's the commercial versus private LLMs. And like I said, there's a lot of corporate data that companies for whatever reason, right or wrong, don't want those data points outside the firewall. So, how do we go about building LLMs internally. So, this is kind of bringing a rise to another platform, platforms for helping companies derive and build these things, because this isn't easy stuff. You need all these GPUs with all these cores. You need to train all these models.

And I think if you've been to any cloud computing conference lately, they're turning into AI conferences. And, really, what they're doing is they're taking the rocket science out of this and really abstracting away all the real, real hard work it takes to train these models and making things much easier for companies to consume. However, they're public cloud computing companies, so that's not going to work for every use case. I think a lot of companies will have some things that they don't have a problem being in the public cloud, but there's other things that they're going to want on their own premise.

So, there's already a race to being the platform winners for that. There's new companies coming in. Of course, all the existing legacy companies are putting that label on their product. But it's an interesting development going on there.

And then, like the marketplace concept in cloud where people could create a service and market it, I expect to see that same thing with LLMs where let's say you're in insurance, or maybe you're in marketing and there may be a model for the use case you're trying to sell already out there in the marketplace that you can go grab and pay for as a service, as opposed to having to build all this stuff yourself.

So, that's a great thing that's emerging as well. And the chips are getting faster. They're driving the cost down. Right now, it's still a little bit of rocket science but I think these are areas you're going to see massive improvements. And of course, I haven't talked about any of the negatives yet, but there's – with every new technology there's the infancy of security and regulatory issues. So, you're going to see a lot of work in that space too.

And Then, the last topic. I'm sure people might be a little tired of hearing about this one, but we keep hearing, "Oh, I'm a developer. I'm going to lose my job." And if I had a dollar for every time I heard this for every technology shift, I remember when cloud came out. I was very early cloud. I mean, when I started in cloud Azure didn't exist; Google didn't exist. It was Amazon, it was GoGrid and Rackspace. That was it. And same thing, "I'm going to lose my job. I right-click and create five servers." Well, guess what happened. The infrastructure people who learned and embraced cloud are very valuable today. And I remember having a podcast, one of my first podcasts at Deloitte. I've been at Deloitte six years now. Gene Kim was on, and he made this statement: "There's never been a better time to be in infrastructure."

And this was six years ago. That's the same time where a lot of infrastructure people in the early cloud days, it wasn't early-early but it was early for a lot of people, were like, "I'm going to lose my job with this cloud thing." And here you go, you've got one of the largest thought leaders, author of many books out there saying, "This is the best time ever to be in infrastructure." Because you can move faster. I mean, it's not as much physical infrastructure; it's more software infrastructure. But when you learn that stuff, what you can produce, you can really be a lot more efficient a lot faster.

And the same thing's happening here with AI. We're starting to make it easy to automate tasks that we probably shouldn't be doing anyway. They're tasks that have got to be done, but they're not high value-add, but they've got to be done. And early AI, where we are today, that's where a lot of low-hanging fruit is and there's a lot of stuff, a lot of tasks being automated using AI. In the software development lifecycle, same thing. You'll have a meeting with the product owner, and you collect all these requirements, and someone's got to write all that down and you've got to create all the user stories and do all this work. It's very valuable work but it's easily replaced by AI. What if the person gathering those requirements didn't have to write all these cases, didn't have to create all the documentation, what if they just listened to the client, asked the questions, and then that audio or that image was captured and those things were created, and then that person could just maybe get 80, 90 percent of the way there, then they could do whatever adjustments needed to be made and it's a huge time save. So, those types of things.

And that's where we're at today on a lot of things. But what's really happening is more large business use cases. You get into health care in some of these places, there's a lot of data, a lot of complexity and ability for the machines to learn and give us some unknown insights and do those types of things. So, a lot of focus today is on the smaller stuff, but in 2030 where we have 100 percent budget in IT on this stuff, you're talking about monumental use cases, new business models. You're talking about more substantial stuff.

And I've been around a while. It's a lot like cloud. When I first started in cloud it was set up a server, and you look where we are today and there's APIs for satellites, retrieving satellite disks. There's APIs for doing machine learning AI. I mean, it's not about the foundation. The foundation is important; it sets the stage. It's about what are we going to build on top of these foundations going forward.

So, today is early. There's a lot of people worried about Dev jobs. Take me, for example, someone who grew up coding an assembler in COBOL and up to C, and all the way up to Python and those types of things. I don't code a whole bunch anymore. So, when ChatGPT got real popular a couple years ago, I went out there and tested it and I was creating applications, not production-ready, but I was creating applications really fast in languages I never used before. And it was really refreshing for me because I got a little rusty from being away from it for a couple years.

So, really cool stuff. Give us some feedback on this. I hope you enjoyed talking about Generative AI. Give us some topics on Generative AI. Give us some subtopics on it, where you want to go, and we'll do another one of these. But thanks for listening today to this Knowledge Short, our first of many to come on Generative AI. Make sure you like us, leave us a review, and subscribe. You can also check out our past episodes wherever you listen to your favorite podcast. You can always find me on Twitter (X), @madgreek65, or reach out to me directly at mkavis@Deloitte.com. Feel free to write me any questions we can address in future episodes. Thanks for listening and see you next time.

Operator:

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