Deloitte.

For Cloud Professionals
August 2021

For Cloud Professionals, part of the On Cloud Podcast

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

Title: **AI/ML: easier, faster, and more powerful with cloud**

Description: AI/ML has been around for decades, but cloud has made it possible for companies to leverage AI/ML in new ways, and on a scale never before possible. In this podcast, David Linthicum talks with Deloitte's Brijesh Singh about why AI/ML is easier to implement with cloud. The pair also discuss new uses for AI/ML—especially vis-à-vis customer service applications—and what's in store for the future of AI/ML. Finally, Brijesh gives advice to practitioners who want to acquire AI/ML skills.

Duration: **00:24:06**

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. Welcome to On Cloud, the podcast for cloud professionals, where we break down the state of cloud computing today and how you can unleash the power of cloud for your enterprise. Now here is your host David Linthicum.

David Linthicum:
Welcome back to the On Cloud podcast, your one place to find out how to make cloud computing work for your enterprise. This is an objective discussion with industry thought leaders who provide their own unique perspective around the pragmatic use of cloud-based technology. Today on the show, we are joined by Brijesh Singh. Brijesh is a consulting member of the technical staff at Deloitte. So, introduce yourself. Tell us what you do, and we're going to talk about a specific project been working on, but also just generally I know we always spin a lot of plates at Deloitte, so what plates are you spinning right now?

Brijesh Singh:
Hello, David. Thank you for having me on the call. I wanted to share with the group that I am a member of Deloitte's AI-based initiative in the “Age of With”. I work in strategy and analytics offering portfolio, and for the past several years, my focus has been on AI-enabled business transformation. Since 2016, I've delivered various engagements for our strategic and platinum clients where we have transformed their multitude of businesses leveraging many tools on the AI spectrum.

David Linthicum:
So, there's a lot of discussion around AI and the ability for it to provide automation where automation wasn't there before. AI's been around for a long time. I'm 58 years old. I worked on AI systems when I was in college, Lisp systems, things like that. So, what has primarily changed in the world of AI and machine learning since then, and why is it such an attractive option now, and why are we seeing it much more these days?

Brijesh Singh:
That's a very good question. I would compartmentalize it into a few domains. One domain is cloud. How cloud is—because of the elastic nature of cloud, what is it that it was offering in the past I would say about eight to ten years where you can get really, really strong compute capability and availability of storage resources at the drop of a hat, within minutes. So, that availability has become easier. Machine learning models that require these compute capabilities now can be executed in a much quicker, faster way. Second thing is that in the past decade, decade and a half, many Fortune 500 organizations have moved their data, or are in the process of also moving their data, from their on prem environment to the cloud environment. They have been building these data warehouses, per se, but going in the environment in the cloud where now data is more easily accessible by data scientists, going on platforms where the exploratory data analysis can be done fairly easily, and then modeling; pre-built models are available through open-source platforms where you can leverage, and you can apply that in a very quick manner.

So, proof of concepts can be executed fairly quickly, demonstrate the capability that that model has offered, and then the ability to take that model and employ it in a production environment, which is now what we are calling MLOps, which is another spin on the word DevOps. How do I conceptualize the model? Where do I pick my data? How do I build my data pipes? How do I—once I have done my exploratory data analysis, how do I now take the data and divide into multiple segments to first train my model, then I test my model, and I find out, based on case course and various factors, which of the model is working well. Then I take that model in the production environment, and I'm continuously measuring the output the model is producing, and I'm monitoring the model drift, and then I refresh my model. This entire cycle, which is now known as MLOps, is being enabled in the cloud environment in a much simpler way. That's the reason why, of late, we are seeing that almost everyone is trying to take advantage of them.

David Linthicum:
Yeah, and the big news is it's really cost-effective now. I remember building these big AI systems—we call it machine learning; then just called it AI. It would suck up $1 million worth of compute cycles leveraging a time share service. Just do some very important simulations. It paid for itself in terms of the value that came in. But now we have kind of weaponized AI/ML for people to allow—I mean, for businesses and organizations, allowing them to punch above their weight. Is that kind of what you're seeing?

Brijesh Singh:
Absolutely. I'll give you another example where some of these deep learning models that are developed by the bigger players like the Googles and Amazons of the world. Now everybody's using those deep learning models, especially in the domain of natural language understanding and natural language processing, where the human-to-human conversation that traditionally have been happening, as any organization—let's take any organization, there are three primary entities that an organization engages with: their customers, their business partners, and their own employees. The traditional way of engaging these three groups was either via a call center, you provide an 800 number and people call in and talk to you. Or they will send e-mails and they will process the e-mails. Then few years back, they started hitting the chat bots are being built where, because of these ML capabilities.

But with the enhancements in speech to text and text to speech, which is now neural, where the voices that respond to you are sounding more humanlike, they're behaving more humanlike because they're supported by the AI models. This transformation of that business engagement that was happening between two humans, now it is happening between a human and a machine. And the ones that are still—okay, so there are two things. One, it is moving in the direction where it's happening between a human and a machine. And the second is, for the remainder that is still happening between a human and a human, the humans are getting augmented. They're getting supported with various AI techniques that are coming in, automating in the processes. So, that's what I'm seeing. We have executed certain programs for some of our clients where we have seen these benefits brought to them in real life.

David Linthicum:
Yeah, and you've been working on conversational AI, which is getting to automate things such as call center interactions and understanding the technology, getting to a point where it's so good that humans are almost preferring to leverage these kinds of conversational AI capabilities, these systems than actually waiting in a queue to talk to a human. Can you tell me a bit more about that?

Brijesh Singh:
Well, yes, David. So, if we look at the marketplace, in the contact center, which was early in the call center world, because the phone calls would come in, what was the traditional way of operating it? A call would come in, there'll be a PBX, then there'll be some kind of an IVR engine that would try to do some call triage and then would—then some routing would come in, and then the routing would take the call and route the call through some sort of an ACD queue to an agent who has a desktop where the soft phone is there and it pops up, and then the agent is the one that is conversing with the caller on the other end and leveraging a plethora of applications, whether CRM applications, back office billing applications and so on, and assisting the person who's on the other end of the line.

Now in what you and I are experiencing, David, and the millennial generation is experiencing that in our daily lives, we are seeing the newer and newer technologies helping us. “Erica” is coming from Bank of America. You have Siri, you have in Google Home where you call in, you have Alexa when you call in and talk. So, since we are getting accustomed to that kind of technology, we tend to then expect the same thing in our business relationships. So, when I'm talking to my utility provider, when I'm talking to my telecommunications service provider, when I'm talking to my bank, when I'm talking to my retailer, I'm expecting that. And that's where conversational AI is becoming more and more mainstream. And what I said earlier, ten years ago, technology wasn't as mature as it is today. There are three aspects of the technology. One is the front end, which is when somebody speaks something, you have to understand

what that person is saying. And for you to understand, first you have to convert that audio part into a text bot. And then you have to take the text bot and apply it to your NLU/NLP engine to map it to a cogent response. That is known as Artemis intent mapping. But prior to that, speech to text is happening.

Ten years ago, technology that was available to us could not do a more robust conversion of speech to text. Now it is happening. Now it is in the high 90s with the confidence score being high. Then the mapping, the classification that is being done of identifying an (Inaudible) and mapping to the right intent, that has tremendously improved. Now, the end part is that coming back with a cogent response in a voice, if you all remember, I think it was about three or three and a half years ago when Google's CEO Sundar Pichai did a demo where he made a restaurant reservation and a barber shop reservation. One could not differentiate the voice, which was the voice of a robot, not a natural human voice. And that's where the technology is, that the neural generated voice can also create personalized voice, and the personalized voice can be created if I want—if I'm making a phone call, and let's say I'm calling my hospital system. And my hospital system chooses to have a custom voice in, say, Taylor Swift's voice.

They have the ability to do that, and I will hear Taylor Swift's voice on the other end greeting me and saying, "Hello, Brijesh, how can I help you? Can I make a reservation with your internal medicine service provider?" So, that's where we are, David, and that's the evolution that is happening in the conversational AI domain. So, in the call center world, in the contact center world, folks are trying to address three main things: customer satisfaction, net promoter score, and then the trustworthiness. To address these three things, technologies are enabling you to do what conversational AI offers to you.

David Linthicum:
So, moving forward, this technology is getting good, and I've actually interacted with a few of those systems, both kind of in a technological demo circumstance, and even in real life. So, are we moving to an area where human-to-human interaction, which I view as kind of inefficient as a technology geek, is going to not necessarily fall by the wayside entirely, but it's going to be reduced to just certain circumstances where the chat bots are not optimal for dealing with that stuff when you have a more higher-level conversation? Because, I guess when you get right down to it, they can deal with 90 percent of the interactions with customers, do so in a more efficient way, and even do so at a higher customer-satisfaction rate, which is something that we didn't anticipate ten years ago when chat bots started to emerge and we had IVR systems, things like that. People were not happy with them, typically hang up on them, things like that. We got an automated voice response, they would move on, or everybody presses zero. That may be changing soon. What are your thoughts on that?

Brijesh Singh:
You're 100 percent right, and I would go one step further in saying that it has already changed. It has already changed. There are some early adopters. There are pioneers who have leveraged this technology, and they're experiencing what you just explained, David. They're experiencing that their digital omni channel, multimodal platform where no human is in the loop is able to contain call containment—by the way, call containment—I'm digressing a bit, but I want to emphasize on this. There are two ways industry looks at call containment. Folks would say, “Okay, if a call came in and if the digital channel touched it, I consider it contained.” In my definition, that is not containment. Containment is that the caller's purpose has been solved. If I called in to inquire on my account, I satisfactorily completed that transaction. If I called in to make a reservation, I did that using the digital channel with no human in the loop.

So, you're right. As the machine learning is powering the engine in the background, enabling the conversational systems to disambiguate these calls more and more, and increasing the predictability of answers that one needs to provide by creating a knowledge graph in the background, more and more we are seeing higher call containment, higher customer satisfaction because those calls—the agent who is spending about 15 minutes, here you can get the same thing done in about 6 or 7 minutes, and with almost 100 percent assurance, the majority of the time when they get it done, it's going to be done right with 100 percent accuracy. When the system doesn't know, when the system cannot do it for you and cannot answer the question for you, it does pass on that conversation to a supervisor, which in this case, could be a human agent saying, "Hey, let me connect you to a supervisor who can assist you further."

David Linthicum:
So, the folks listening to the podcasts are mostly cloud technologists. They know this technology exists, but perhaps didn't know that it was at the level of sophistication the way you're describing. I'm finding out about it recently as well, so what should they be considering? Folks who are building systems for businesses, when these things should be implemented? Because I suspect that they're probably going to become more popular as time moves on. It's going to be a key tool to make it happen, but what technology should they be looking at right now in terms of the technology categories, not particular brands, and where should they be looking to find out more around this technology and how it's working? Other than listening to this podcast, which is a good first step.

Brijesh Singh:
That's a great question. So, if I—again, there are two parts, David. One is the machine learning, model building part of it, and then there's this whole conversational AI part of it, so I'm augmenting the conversation, human-to-human conversation, or improving and making it better. So, of course the natural language understanding and natural language processing part of it with deep learning, which came about 30, 35 years ago, has evolved through the process and has attained a status where it is so accurate today that if you don't plan to use it, you're falling behind because your competitor is using it. And then again, the natural language generation is not there yet, in my humble opinion. That is one area where organizations are trying to get to that state, and I'm hoping that in the next couple of years, we will be there. The text-to-speech part of it where the neural part is coming, the neural TPS technology has evolved so much that what Sundar Pichai presented three years ago has become more a reality now. I won't say that it's there yet. The voice is still not 100 percent human, but it is almost there, and you can customize it, and you can make it a human voice, should you prefer to do so.

On the model side, on the pure-play building data science generating insight, I would say there are various platforms that are available in the cloud environment where they can plug in with your cloud service provider, and if you have brought in your data and just storing your data in a standard, unstructured format, some of these newer players can get that data and present that data in an easily-accessible fashion where your data scientists now can develop those models much more quickly and employ those models much more quickly, generating insights, advising information to you that you can take corrective business actions on. So, I'm not pointing out specific technology per se, but I'm calling out that these—I want to call them data lakes, but I will call them cloud-enabled machine learning tools and techniques. Those are available in the marketplace from various providers that one should take advantage of.

David Linthicum:

So, what are the current limitations that need to be overcome with this technology? Obviously, it takes a tremendous amount of processing. It's newer technology, so I assume the complexity is there, and I guess it's probably tough to find people who are skilled in this various technology. What would you warn people about? There's an upside in leveraging this technology, but there's always a downside. What are the downsides that people need to consider?

Brijesh Singh:
So, I would not call it a downside, David. I would call it my lessons learned and things to be aware of and plan for. So, integration. Because, at the end of the day, it's a complete governance process where you are consuming data, and the data consumption is not a one-day event or a one-stop event. It's a continuous process. So, how do you consume the right data, the right amount of data at the right time? And then you continue to do so. Then how are you integrating it with your other systems? How are you ensuring that this integration is secure? You're following the compliance and regulatory requirements, especially if you are in certain industries like healthcare or financial services. These are some of the areas that I would encourage my listeners—our listeners, rather, to be aware of and plan around.

David Linthicum:
So, moving forward, where can the skills be obtained for this? In other words, it's certainly something that's been around for a while, but in taking this thing to the next level, I'm assuming there's another layer of technology that we need to kind of understand. So, say you're staring out. Let's say you're a cloud architect and you're moving toward this particular technology as sort of an overall skill set. So, where do you go to learn things? What kind of courses should you take? Are there courses out there? Where do you go to read information up on it? And also, would it be something worth pursuing as a majority of your career, not only just AI, but the ability to kind of deal with AI chat bot technology as related to conversational AI?

Brijesh Singh:
Well, as you and I both know that if you are in the world of computer science as a software developer, as a coder, then it's an evolving process. You can't just say that what I learned in school and college, that one or two languages that I learned, or one or two platforms that I learned, I'm going to continue to work the rest of my life on the same platform, with the same languages. I have learned from 3GL to 4GL and continuing forward from a relational database to NoSQL, and continue forward in a graph database, so that's the evolution. What I would suggest our listeners, especially younger listeners, to do is that understand the basic concept of software development, full-stack development cycle, full-stack development process you should have a knowledge of. You should have a knowledge of the DevOps cycle to learn about DevOps. Learn about maintaining code integrity. Learn about quality check and quality control on code. How do you do code review? Where do you code review?

Learn about the testing process, especially in the conversational AI domain. And you asked me this question earlier. I remember this, that one of the guardrails to care for, especially testing for a voice-enabled conversational system, is testing. Because this is a very, very different domain. It is not a typical software development cycle. You have to worry about usability testing. You have to worry about use of experience that comes into play, different accents come into play, different regions, people with different ethnicities speaking different accents, different choice of phrases and words. So, you have to test for those. You have to test for ambiguity that there could be background noise coming in. So, there are certain special considerations that need to be made.

So, from a tech point of view, the earlier part that I was saying that try to be a full-stack developer, try to learn some of these newer languages like Python, Java. If you're familiar with that, learn about containerization when you are in the cloud. Learn about Kubernetes, what does that mean. And it depends on the cloud service providers. If you go on from cloud service provider A to B, some of the parameters will change that, maybe if you create a containerized software on cloud provider A's platform, it will time out in, let's say, ten minutes. On provider B's platform, it will time out in 15 minutes. Or you can increase the capacity. So, you have as a solution architect for a large service provider, you would learn that, so I would say go and get certified as a solution architect for top two or three or four cloud service providers. Learn about what hybrid cloud means. Some folks are advocating the, “big edge hybrid cloud.” What does that mean? Learn about that. So, these are some of the things that I would advocate folks to learn.

David Linthicum:
I think that's great advice, and it's a good way to leave the podcast. Anyway, appreciate you being on the show, and what a fascinating technology this is, and it's great to watch it evolve and how it's going to make our lives better.

So, if you enjoyed this podcast, make sure to like and subscribe on iTunes or wherever you get your podcasts. Also don't forget to rate us. Also check out our past episodes including the On Cloud podcast hosted by my good friend, Mike Kavis, and his show and book by the same name, "Architecting the Cloud." If you'd like to learn more about Deloitte's cloud capabilities, check out deloittecloudpodcast.com. If you'd like to contact me directly, you can reach me at dlinthicum@deloitte.com. So, until next time, best of luck with your cloud projects. We'll talk again real soon. You guys stay safe and have a good summer.

Operator:
Thank you for listening to On Cloud for Cloud Professionals with David Linthicum. Connect with David on Twitter and LinkedIn and visit the Deloitte On Cloud blog at www.deloitte.com/us/deloitte-on-cloud-blog. Be sure to rate and review the show on your favorite podcast app.

About Deloitte
