Deloitte.

The Deloitte On Cloud Podcast

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

Title: **AI and cloud are delivering higher quality, more actionable information**

Description: Artificial Intelligence (AI) has been around for 40-plus years, but its future has never been brighter. In this episode, David Linthicum talks with Professor Vasudeva Varma and Deloitte's Ponnu Kailasam about how AI, in the context of information retrieval, can wade through information overload in search results to deliver contextually-appropriate, actionable information. Vasu and Ponnu agree that AI powered by cloud, and cloud powered by AI, will be critical to the process.

Duration: **00:24:55**

David Linthicum:
Welcome back to the On Cloud Podcast. Today on the show I am joined by Vasudeva—Vasu for short—Varma, and he's a professor at International Institute of Information Technology, Hyderabad, which is a city I've visited—love it there—currently serving as the head of Language Technology Research Center. And Ponnu Kailasam—hopefully I didn't destroy that too much, Ponnu—he's a managing director at Deloitte, a colleague of mine, and leader of the cloud engineering practice in India. But let me get into this, because this is an incredibly interesting set of backgrounds that both of you have, but specifically

Vasu. So, Vasu, kind of tell me how you got into computers and your kind of a quick background into your education and your motivations, and also what you do in your day job right now.

Vasudeva Varma:
Great. First of all, thank you, Dave, for the opportunity. I'm delighted to be here. It's an honor to be with you and Ponnu. I'm a professor at this place called IIIT, I-I-I-T Hyderabad, which is a university focused on information technology. So, I work in the areas that are related to text processing, NLP, social media analytics, and right now I teach courses like information retrieval (IR), NLP, AI, ML, social computing, and so on. And, well, I got into academics after kind of spending a significant amount of time in the industry. I worked in New York for some time and then in the Bay Area for about five, six years. When I moved back to India, I stayed—went to academics because that's where my heart is. And I started experimenting with bringing in some of the best practices of learning and doing from the industry, and it's been a wonderful journey ever since.

David Linthicum:
So, Ponnu, kind of same question. I mean, what do you do during your day job? I love talking to colleagues I have at Deloitte because it's really the only time that we have, because we're so busy, to learn who we're working with. So, kind of give us your story.

Ponnu Kailasam:
Sure. Thank you, David, for having me, and it's an honor to be in conversation with you and Professor Vasu. I'm part of the Deloitte Bangalore office, and two days from now I actually complete 19 years in Deloitte. Pretty much most of my work experience has been with Deloitte. I've been very technology-focused, and currently I play two roles. One is I lead cloud-engineering practice for Deloitte in India, and I also play the co-dean for Cloud Institute for Deloitte. My day-to-day, my focus primarily is to build new capabilities and build the cloud engineering practice for Deloitte.

David Linthicum:
Wow, 20 years. That's actually pretty common. What I hear, working at Deloitte. In other words, I've been here about five years and I'm kind of a baby compared to everybody else. Everybody's been here for such a long time. So, Professor Vasu, let's go ahead and get to you. How did you enter into cloud computing? What was kind of the motivation and what kind of technology patterns interested you in kind of moving into the cloud space?

Vasudeva Varma:
My cloud computing journey started I guess 15 to 17 years ago in 2005, 2007. I was a hardcore IR researcher. I hope to believe I still am. But then at that time we were trying to build a search engine for all Indian languages, not just a toy engine for demo or academic focuses, but a real one that works, right, the one that works better than any other search engines at that time. So, we cracked some of the hardest Indian-language processing challenges and we got into creating very effective retrieval and ranking algorithms. So, we wanted to call all Indian language content, and then we hit the wall of we couldn't call all the content—too many issues at that time.

So, that's when it struck me that I need to go beyond my comfort zone, get into the distributed computing and so on. So, of course, we were aware of some of the open-source search engine efforts at that time like Lucene and Nutch. So, in 2006 or 2007, I met this person called Doug Cutting at a conference in Helsinki. So, Doug was famous in our community, in information retrieval community, as the one who created—he invented this famous Lucene and Nutch. So, I went to Doug and told him what some of the issues were that we were facing in making our distributed IR architecture work. So, he told me about a new project that he pitched launched called Hadoop and told me that Hadoop should be able to solve many of the problems that I mentioned.

So, we ended up being one of the early adopters of Hadoop and we started playing with it. We also contributed to its improvement. So, we have built a search node using our regular desktop machines, and we created a near-commercial search engine quality solution for Indian languages. And then we later extended this search solution to include 250 world languages that have kind of having similar properties, like agglutinativeness and so on.

So, at this point in time, as I started enjoying working with the system side and then started looking at the distributed IR as a very interesting area, I have started pitching a module in my information retrieval course on distributed IR, which included topics like MapReduce programming and Hadoop and so on. And then in the next three to four years, this grew from a two-lecture module to almost like one-third of the course.

So, I was thinking that my cloud computing is eating my information-retrieval content, so I started offering—I kind of decoupled them and started offering a full-fledged course in cloud computing. Perhaps we were among the first ones to offer a complete, a full course on cloud computing.

So, I also, at that time, I was not just teaching and using cloud infrastructure. I also saw that some of my students, being more excited about these aspects like distributed and system aspects, and then looking at their excitement, I started developing more deeper interest. So, our cloud computing grew to—kind of expanded from two to three people to like—almost like a dozen people working on some of these problems. And we started addressing some of the interesting research programs like applying ML to cloud tasks like job scheduling and so on.

And then we had kind a heterogenous infrastructure—being an academic institution—we couldn't really afford uniform high-grade nodes, so we had a heterogenous infrastructure. So, we started working on problems like cloud migration and speculative execution within the heterogenous infrastructure. We also had a few papers on these heterogenous clouds and actually predicting the performance, and also SML—monitoring, detecting some of the SLA violations and so on.

So, we had a fantastic, productive time in the core cloud computing area starting from 2010 till something like 2018 or so. So, that's been my kind of the journey with cloud computing.

David Linthicum:
So, Ponnu, what's your reaction to the research that Vasu's doing in terms of how it kind of incorporates into our theme of building cloud solutions for clients?

Ponnu Kailasam:

Yeah, absolutely, because as cloud evolves and also converges with artificial intelligence. I think that natural language processing, the ability to actually understand different languages, different backgrounds is going to be I think very, very critical to that.

David Linthicum:
Yeah, I think so. So, Vasu, I've got one for you. I'm a reformed college professor myself, so what do you like better, teaching or research?

Vasudeva Varma:
Oh. Can we have one without the other? I really doubt it because one fuels the other. I don't think I can just stay purely on the research side or on the teaching side, because like it's actually the tiny bit of research that we get the students to do during their course project or something that's very exciting for me. So, I can't really choose. I need both.

David Linthicum:
I need both, too. I agree. You can't have one without the other. So, cognitive or semantic search and research really kind of is the topic. This is really what we're talking about here, so the ability to not only deal with requests that are coming from people who are looking to find information and topics and perhaps even binary data—videos, audio, things like that—on the web, but the ability to find it in an accurate and targeted way. Someone who does a ton of research on the web this is an invaluable tool. So, how has your research kind of bore out and what's been the evolution in your thinking in terms of how we do semantic search? And what are some of the things that folks who are looking to get into this area, focus area of research should understand?

Vasudeva Varma:
Oh, thank you for that question, Dave, because it's very close to my heart. What I think is retrieving information is a solved problem today. I think we know—we have the knowhow of retrieving a piece of information, no matter how deep it exists in the content that you have. But then the real problem is information overload, so the kind of information that's relevant and that's available to us, and how do you really organize this? And how do you really make it available to people when they need it, in the form that they need it, and then also in a way that they can do something with it in their task? In other words, the semantics of a cognitive search should be invisible, and it should actually—its effectiveness should come when we have this moment of the need, right?

So, if you think about information overload as the problem, there are a bunch of ways to solve it. And, of course, summarization is one of the key solutions. So, wherever is the information across sources, kind of getting this relevant information, then organizing it, and then presenting it to the users at their time of need is a very important area. So, we've been working in the areas of multi-document, clearly-focused, abstractive summarization areas, and that within various flavors of summarization. And then like the summarization should be really the solution for most of the information overload problems, right?

Some other aspects of the semantic search are really, like intent detection, rather than the surface-level keyword search. So, what do, really, the users want? Why are they performing the search? Can we get to the intent level and then provide the information that they do? In other words, we need something like two engines, not just the search engines, so—and we need results that are actionable. And in the last decade or so, a lot of work that went into creating these actionable results, and then we have much better search engines today, but then I think we have only touched the tip of the iceberg. So, we have lots of other things to uncover, especially when we get into verticals such as domain-specific searches and so on.

So—and another important thing perhaps is the context of Web 3.0. We will see a lot of content and information being produced by the machines, not by humans alone. So, we need to really reimagine our search engine technologies to address this massively proliferated information overload. So, in other words, they need to be more intelligent, should be able to really understand the semantics of not just human-generated content and the questions, but also the context in which the machines are generating content.

So—and another key element that is related to semantic search is the ways in which we can repurpose the good quality content. So, for example, there's a good technical paper, but then we wanted to really communicate some of the important aspects of that technical paper to, say, a layman. So, how do you really do the style transfer from formal to informal, from heavily-technical to lightweight content? So, the style transfer aspects of the search—that is basically a layer on top of cognitive and semantic search, these are all the things I believe from the core of cognitive and semantic search area, and that's what I'm really excited about, all the possibilities in this area.

David Linthicum:
So, I love the idea that you have—you're building systems that are able to determine intent, which is something—I may not even know my intent in looking for the information. So, what are kind of the mechanisms around doing that in the back end? Are you able to do it with some sort of degree of accuracy, or is this basically a learned response where you get better over time?

Vasudeva Varma:
Absolutely. So, I know, of course, one key aspect is the language, right, language—what does the language convey? For example, there are lots of pragmatics that we use. If I ask you, "Do you know what time it is?" And then practically speaking, if you say yes or no, then that's not the answer I'm expecting. I didn't ask you for the exact time now. But my intent is to know the time, but then I didn't really pose my query in that way, right? And when you extend that to all other things where—how we communicate with people, so the pragmatics is an important aspect.

And then there are the language semantics-driven solutions, but in addition to that, the deep-learning era that we are in currently, it is often in completely new kinds of solutions where, somehow, you're able to capture the user's behavior, the user the reactions for a given query and a bunch of results that come in. So, the systems are able to predict the intent somewhat automatically. So, lots of data, through it these machine learning, deep learning engines, and then the automatic detection of intent—this is a fair amount of success in this particular area. And I'm really—I'm sure that we'll be able to see more of it in the very near future.

David Linthicum:
Yeah, it's amazing technology to me. It comes to the productivity, and you think about it. We're learning finally—and, by the way, I'm almost 60 years old and I was an AI analyst as my first job out of college when I was 20 years old. But it's evolved in such a way where it's becoming finally productive, our ability

to do it, but, also, it's finally affordable. In other words, cloud computing, even though it has brought the ability to deal with this stuff in an economical way—we're able to consume AI services, but also consume AI services that are head and shoulders advanced of where they were just even ten years ago.

So, moving forward—and, Vasu, I'm going to go to you first on this one. So, how will AI be evolving? And what are some of the things that we should be looking for and additional capabilities that'll start to show up shortly and perhaps show up long term?

Vasudeva Varma:
Right. So, maybe let me start with a little bit of a background here. I think we are in very interesting times. Somebody said that we are in the third connected age. And the first connected age, which is somewhere in the mid-'90s to mid-2000s where we were actually connecting with each other with machines to discover information and to reconnect through these web services, to connect our transactions and that's where ecommerce and search engines flourished.

Then mid-2000 to somewhere in the mid someplace 2015 or so, there is a second connected age where the social media proliferated, and mobile-first applications have come in. And then everybody is connected to everybody else, and then we are also connected all the time, right? So, this is the kind of new characteristics of the second connected age.

And the third connected age that we are in, right—so we want really faster connections, like five-plus Gs will enable all of that so much faster and kind of resilient forms of connection. So, the latency is no longer an issue in most cases, right? And then, of course, with all the new ways of connecting, with voice, with augmented reality, virtual reality, there are a bunch of things that we are kind of experiencing for the first time, and with a bunch of new technologies, including AI like blockchain.

So, we have really an entire new world of Web 3.0 metaverse, tokens, wallets whatever that you like to kind of name it here. So, we're just they're all, like—all those are the kind of part of the changes that we see that are on the way in this third connected age. So, the key is that every individual and perhaps every organization will have to pay close attention to the implications and opportunities of this third connected age.

So, I wanted to kind of get to this background before really talking about the role of AI and so on, right? So, see—some of the interesting things that we are doing within this background especially, where we have OpenAI systems like GPT-3 and recently DALL-E where machines are generating interesting content, machines are generating interesting visual presentations and so on, but the trust aspect and then making sure that this generation is not—that the content generation is not really hallucinating there are—it is grounded and then it's actually based on something constant and those aspects, right? So, we need to really look at the data-to-data connectivity when you're looking at the cloud and AI together.

So, for me, one of the important aspects are twofold when you are trying to connect cloud and AI. One is that AI in cloud, for example—AI-as-a-service, so most of the cloud providers today are offering a bunch of AI-as-a-service. And then other aspect is AI for cloud, so for example, AIOps, in other words, seeing how—how significant they're becoming, how important they are now, and where the mission-assisted human operations or human-assisted mission operations, the spectrum is kind of becoming very, very interesting.

So, in short, I just would like to say that we have witnessed a very important tipping point where AI can actually influence the cloud in terms of the devices in our hands, to sense us, to intelligence at the edge, intelligence at the cloud's edge, and intelligence in the cloud, right? And then I think the possibilities are endless.

David Linthicum:
Absolutely. So, Ponnu, bring us home—last question. So, what are your thoughts on this as far as the destination of this technology? And what's your thinking around how this technology will be evolving?

Ponnu Kailasam:
Yeah, David. I pretty much completely agree with Professor Vasu. I think that the trajectory for AI and cloud will be an interwoven one, and then when you add edge to it, the 5G and the edge to it, I think that will be the triumvirate of technologies which will be the—it will define the future. Cloud will help adopt and scale AI, so giving a lot of opportunities to organizations and give the benefits of deploying and optimizing the AI deployments for them. And the edge along—combined with that, the edge is going to really help the end users with both the scalability of cloud and the intelligence of AI.

David Linthicum:
Great insights, both of you. I really enjoyed this conversation.
So, 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 my good friend, Mike Kavis. Find out more at DeloitteCloudPodcast.com, all one word. If you'd like to contact me directly, you can e-mail me at DLinthicum@Deloitte.com, L-I-N-T-H-I-C-U-M. So, until next time, best of luck in your cloud journey. Everybody stay safe. Take care.

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