The Deloitte On Cloud Podcast

Gary Arora, Chief Architect, Cloud and AI Solutions, Deloitte Consulting LLP

Title: **Google Cloud Next 25: Deloitte CTOs talk Agents, AI, and the future of Tech**

Description: Live from the Google Cloud Next expo floor, Gary Arora talks with Deloitte's Bill Briggs and Mamoun Hirzalla about the latest announcements from Google Cloud Next '25. They explore how AI agents are revolutionizing business processes and the importance of fine-tuning large language models (LLMs) for specific scenarios. The discussion delves into the integration of AI in legacy systems, the role of consortiums in setting standards, and the future of AI hardware. The trio also discusses the challenges of managing AI agents, the need for interoperability, and how collaborative efforts are shaping the future of AI technology.

Duration: **00:26:11**

Gary Arora

Hey, welcome back to On Cloud. I am Gary Arora, your host, Chief Architect for Cloud and AI Solutions at Deloitte. We are coming to you from our pop-up studio at Google Next conference in Las Vegas. Now, if you've been following the keynote, Google has dropped a lot of announcements on AI, data, and so many other things. To figure out what that means for you and your business, I am joined by two incredible guests, Bill Briggs, the Chief Technology Officer at Deloitte**.** And Mamoun Hirzalla, the Chief Technology Officer for Google at Deloitte.

Gary Arora

So, this is the war of CTOs.

Mamoun Hirzalla

Yes.

Bill Briggs

Alright.

Gary Arora

Alright. Thank you so much for joining.

Bill Briggs

I see from the beginning, Mamoun is the man, but it's always great to see you. What an awesome week. Time to dig in.

Gary Arora

Exactly and let's get started. So, what's jumped to you from all the announcements that happened yesterday. Bill, let's start with you.

Bill Briggs

I think that a year ago, an agent was a whisperer and all the action, all the answers were on the model and had potential, how many tokens do we have. We've really shifted from, of course, the tech is front and center, we'll dig into it, but so many more real client use cases of the technology being applied to real business problems across industries, across the globe, and I think that's what we need to see more of and we see some awesome results already.

Mamoun Hirzalla

Absolutely. I mean, last year, it was all about GenAI. Last year, if you look at it, not much about Google Cloud. Yes, there were some announcements, but everything was GenAI. This year, post the Next from last year, we've been seeing all the things that are coming, and we knew that it's going to be agentic, agentic, agentic – and they did not disappoint. Lots of announcements about agentic and what clients can do as part of that and most importantly, over the last six to nine months, we've had a ton of conversations with our clients to ask what does agentic mean for my business? How does that impact my business model? What are these new processes or things that can benefit from that? We'll talk about that shortly here, but it's really about all being agentic and what agentic can do for the business. That's really the exciting part about it.

Gary Arora

Exactly. What excites me the most is that the kind of problems that we can now solve, and they have plagued these organizations for decades, in terms of data silos and disjointed workflows and unlabeled, unstructured data. Now, you have new tools and new processes to actually tackle those issues. Can you share some client examples, client stories where we are actually doing this?

Mamoun Hirzalla

Sure. We have multiple clients. I'll share maybe two or three of them. A large healthcare company right now, we're talking very seriously to them about looking an existing process that has a lot of pain points related to looking at claims data and figuring out what is the right adjudication of that particular claim, how to sift through a ton of information, structured and unstructured, and then that amount of time it takes the current process to figure out what is the right thing to do and then now suddenly looking at it and say multiple disparate data sources, can I go and link them, look across all these data sources without costing me an arm and a leg and then figuring out the relationships between these tidbits of information across multiple data repositories and then coming back with a decision and once you make that decision, have the status update into the backend integrated system through an API interface and all of that, all being wrapped up within the auspices of an agent that's basically doing that. So, this is a very powerful type of use case that clients are looking into.

Bill Briggs

Look, I've been doing this for a long time. So, automation in health care and insurance has been one of those things we've seen investment over investment over in automation, how do we get advanced analytics, how do we get advanced tech. That's been a journey we've been on for a long time, but it always still came down to a whole lot of human reasoning at work. That's the point now where we're seeing. We're just able to go and do complex tasks and every claim and every case is different. So, having to go through and come back with grounded reasoned conclusions. That's the piece that you would always forego before. We're seeing a different kind of return, which is amazing.

Mamoun Hirzalla

I'm glad that you said that because you know many of our clients come to us with, OK, what I mentioned here, by the way, something we used to do. So, what's different? What's agentic about it? So, the difference is the LLM and it's not just the generic LLM that Google releases. We look at the data, we look at the patterns of the data. We finetune the LLM to say, for these type of scenarios, you annotate it, you train it, and say, when you encounter this, this is what it means. So, that's the beauty of it. That part wasn't there a year ago or two years ago. It was just a pure simple API integration. The integration is still there, but the reasoning that happens before all of that to reach a very nice conclusion. Yes or no, it doesn't matter, but at the end, you're saving a human agent from going through that process and giving them a ton of knowledge based on the training that we've done to come back with a conclusion, and that's extremely powerful.

Gary Arora

Exactly. You mentioned the example of claim adjudication system. That's the heart of business for any healthcare organization. For decades, these processes have lived in mainframe. While mainframe is a workhorse, there is a bit of inflexibility that comes with it. What can you do with this data? What can you do with these processes?

Bill Briggs

Something Deloitte does really well, we try to look at it, but, of course, we're deeply curious about the ideas in the headlines, and we're helping shape a lot of the technology that gets released. We've been working closely with it for a long time. So, it's great to see it coming out part of some of the things we've been working on for a while, but we also say here's what else has to be true, including in the legacy application stack, like how does that have to be ready to participate in the future. How do you think about infrastructure hardware, data security, the things that CIOs and CEOs have been around the block? They'll see the headlines and roll their eyes and be skeptical.

Mamoun Hirzalla

Another hype cycle.

Bill Briggs

Totally. We've seen this before and if you only focus on the shiny object of the new, how to make it real, what has to be your best investment? I think the thing we help is to tell that story to CEOs, CFOs, and boards that it's not you have to eat your vegetables before you can have your dessert. We're not trying. It's that there's real advances happening, but there's real investment needed, which, by the way, is what's dominating most of our clients' technology stack.

Mamoun Hirzalla

I'm going to age myself now a little bit. I remember the old days when Java came up, then Enterprise JavaBeans came up, and then Web services. Everybody who had a code in Java became an EJB unnecessarily because they were triggered happy on EJBs, then the same thing became Web services, microservices, and now the same thing we're seeing the pattern, agents. Everybody who has a microservice out there is saying, oh, this can be an agent, but is that the right way to think about it? No, but look, this hype worked out in the past. You go through this process of about a year, you sort out what works, what doesn't, and then it settles on what is the right balance and that's where this will kick off. So, next year, it's going to be a blockbuster year because this is hype, we'll settle down in about, I would say two, three, four, five months and then the true things will emerge as part of it.

Bill Briggs

If anyone had on their bingo card, EJB was going to come out. I'm going to double down now because a huge scene of next year is on interoperability. So, as a guy that helped build and lead our integration practice for a while. If you come to core, but we're going to throw a couple of things there for you all to take, but there you go. Now, we have A2A MCP and the idea of the promise of our AI future requires not just agents who talk to themselves, but to be able to deeply integrated in the business process and applications. So, that was a missing link for a long time. I think that might be one of the things we look back on and say what's different now and we were here first.

Mamoun Hirzalla

Excellent point. That's exactly the discussion I had with a large European bank yesterday because we were talking about, great, everybody's doing agents now. I am doing this process in antimoney laundering, 20 other banks are doing the same thing. It doesn't make sense to do it. Why can't somebody like Deloitte or others who have the industry knowledge that partner with banks make this and make it available. Great. That's the same question came up when CORBA came up, when Web services came up and we said, BPMN will save the day with the agent, with the processes and composable business services never took anywhere. Now, what's different this time is that required a ton of understanding, required a ton of collaboration around the protocols that did not happen. Everybody wanted to do their thing. The LLM is the biggest differentiator now. The agent-to-agent announcement that was happening or happened with Google Cloud, which we were part of the initial discussions as part of the consortium with additional 60-plus companies that Google announced is going to be a significant gamechanger because nobody will trust one company to come and say, I'm going to own the brain trust on agent to agent. Everybody knows this is the next big thing. So, it has to be a consortium lead, which is happening. Then, the third part is how to make the data available to the models. The traditional way is connectors and we help Google build a lot of connectors for their agent space as part of the software engineering effort, but then MCP, what Bill mentioned earlier, is a big differentiator because it's standardizes how you make the data available, the context available, to that model so that as a company I can subscribe or participate or whatever it is to make that data available. So, I'm not going to say think of it as a clean room for models to share, but it's getting close.

Gary Arora

I want to double click on that because I am excited about that update. With agents, I think AI agents are having their microservice moment where when microservices came out and just scaled up, it was this is the first time the service is now exposed for another entity to consume and AI agents are doing the same thing, except now with intelligence on it.

Mamoun Hirzalla

Correct.

Gary Arora

But we also have that notion of everything can be an agent and you may have too many agents and then how do they communicate.

Bill Briggs

What I think about what we learned in the API gateway movement before. Suddenly, it was discoverability, how do we understand the context and the expected outcome in ways that can be defined so the protocols can play, but then the how do we route, how do we put thresholds to manage performance

potentially because we're building it in ways that we can't anticipate how it's going to be used and how it's going to evolve. Which is exciting, but it could be runaway, and let's be honest, it could be runaway costs and complexity. It could be runaway business process operations that we don't understand, you can't explain, which is an issue in regulated industry. So, I think that hard work that we've lived and learned from the integration world will come in handy in this. It requires the investments and the scaffolding around it, or it could be addition by subtraction.

Gary Arora

Exactly. The pace of innovation in this AI agent space has been so staggering that I feel one of the things that was missing in this space was a set of standards, frameworks, the interoperability kind of notion. Do you think that dust is finally settling on some of those aspects with agent-to-agent protocol, MCP?

Bill Briggs

I would say settling not settle, but I think to Mamoun's point about the consortium then, we think about 10 years ago, if you went to a conference like Google Cloud Next, you would not see all the partners. And I like Google speaking on it, they basically said, we're going to continue to invest up and down the stack in tools and capabilities because we can at our scale. Because we know you have to allow multiple participants, because there's going to be so much innovation. So, we're seeing more and more deals with the multiple parties, I think we have to have some opinions of what are the right combinations for a given market industry solution, which is how Deloitte is investing in our own assets and platforms. If you see our announcement with Google Cloud and ServiceNow, around 100 agents that we've created to bring the market, part of Deloitte's evolution from just the services firm that how do we codify our knowledge of industry function into outcome-based software and agents and the likes.

Mamoun Hirzalla

Agree. I think what you touched on is one of the lessons learned from the past. People realize that you cannot do it alone. What they used to do is, in the past, they used to go and try to make these standards, bake them off until they're in a nice state, then release them, and start looking for consortiums. What's happening now is the complete opposite. What basically they're saying, I have an idea, I have four or five pages about it, let's build the consortium about it, build the excitement, and let everybody influence that direction. So, therefore everybody feels that I have a stake in the game. I'm not just following a direction from a big CSP to be part of that. No, I have a stake, and I can influence it, and therefore, it's to my benefit to be part of that consortium and going forward. So, that's the smart move from Google side to be on that front and I think a lot of people trust that consortium of 65-plus companies to be deep-brain trust for it. Google is embracing MCP as part of the process and they're coming with this one and I hope that in the coming few days, we will hear announcements from other CSPs joining the same effort going into it.

Gary Arora

So, coming together of these different entities, stakeholders, to define what those common standards should be, that's the most exciting part because I feel nowhere in history we have had a new technology evolution and setting of standards happened so closely together.

Bill Briggs

I'm going to roll my eyes on to say this, but this year in Vegas a couple months ago. I made the point at that time, we had so many billboards and booths prominently displaying AI is the future. That was the standout. And my my feedback in the moment, was "listen, there's no booth talking about 'now with electrons, or now with electricity.' We need to stop focusing on the artificial AI headline and what it's doing with future set, but with electricity, we don't question how we're getting 110 volts and 20 amps coming out of electricity. You plug it in. It happens. It could be from nuclear, it could be from turbines, it could be from gas. That's the piece that we need to get to where it doesn't mean that hard work doesn't exist, but it's below the line of what most businesses and enterprises are thinking about it. So, then what do we do with it? We're close to it, but it's not past tense soil, but steps in the right direction for sure.

Mamoun Hirzalla

I agree 100%. Then, the other point you touched briefly on it with the proliferation of agents, you mentioned that too, how do you govern all of that? Where is the body that's going to regulate? What's a good agent, what's a bad agent? There is going to be a lot of marketplaces for agents. Google is certainly coming up with one with their announcement, and I'm sure others have similar ones and will announce more. So, who's going to come and say, an agent that checks out these following traits and functions and then somebody needs to rate these agents and how can you be objective and how you take all of that to the market and now how do you discover these agents so that you can build these composable things out of them and trust that this is going to be the key? That area is still very active right now and it's going to be a lot of interesting things coming on that front.

Bill Briggs

So, just saying that every option were out there. Management agents, and middle management agents, and regulatory agents, and executive agents, and agents totaled all the way down. No I'm just kidding, I'm not trying to be flippant. You're exactly right and that is the piece to unfold that we're Deloitte is going to lean in and help share that.

Mamoun Hirzalla

Absolutely.

Gary Arora

No, we don't want to replicate that bureaucracy in AI agents.

Mamoun Hirzalla

No, no, no, no.

Gary Arora

Bill, I liked your analogy of electricity, and I feel what these interoperability adapters are doing is what really your electric adapters are doing. You take your MacBook to any other country and then have a different voltage for it, you put your adapter in, and it does the translation for you. You don't have to figure out how this is going to work. Is it going to blow up your laptop or not? I think that's the moment that we're looking at right now. One of the tech trends that I've noticed this year is that hardware is eating the world. This is so different because it's mainly been about software. So, what is this trend?

Bill Briggs

So, relegated infrastructure and hardware is a commodity that someone is going to figure it out and work that.

Gary Arora

Exactly. So, hardware is having its moment. Tell us what's important about this and how have you seen this unfold with the clients?

Bill Briggs

Even this week with Google's announcements, the reason we wanted to highlight it is the flurry of investment in AI chipsets, server side, and then the suite of announcements from a lot of the end user computer, laptops, and mobile phones of saying we have AI chipsets optimized for the edge, and then specialized hardware that's actually on the edge inside the factory floor or on a hospital floor in a semiautonomous. So, the idea that the hardware is actually opening up new capabilities for the first time and you have to make some decisions, are we going to rent or buy cloud delivery? Is that the right way to go? Quantum is continuing to advance. We've been investing in quantum computing for six-plus years. Should we buy a quantum computer, or should we just think of that as cloud service? So, the idea that it's becoming not just important strategic concern, but it's something more and more CEOs and CFOs are thinking actively about, like where are we going? How do we partner? In what ways? We need to build data centers ourselves. We need to partner with organizations that can provide the capacity. So, that thread has been really important. If you look at what Google announced with their 7th Gen GPU, the Thunderbolt. They have announced to other hardware players. So, they're saying, we're going to continue to provide options for performance, options for power consumption and costs, and be open to allow you to deploy our technology and others on whatever is the right mix for you. Our CEO was talking about 42.5 X slots, which is the maximum configuration of order, which I never thought of it that may hear that, which made me very happy. So, we have this moment in time that we have to understand what it means, like anything. You have to lead with why would we use it before we drive any kind of costly hardware refresh or buyout, but it is absolutely a massive piece with this conversation. If you get back to the CIOs and CTOs being skeptical of the really shallow, empty-calorie headlines and announcements and you start saying what would have to be true and investments have to be in place, that's what makes it real.

Mamoun Hirzalla

Agree 100%.

Gary Arora

For the kind of problems we want to solve, for the kind of innovation we want to drive, it can't just be software alone that is getting optimized. It has to be the hardware itself. So, I'm glad that this is all coming together at such an unprecedented rate. Let me ask you this, we are in the fourth month of this year.

Bill Briggs

By the way, I think I'm influenced by the new Marvel movie that's coming out because my phone buzzed as I said, Thunderbolt, it's Ironwood.

Gary Arora

It's always listening.

Bill Briggs

Thunderbolts is the new Marvel movie, Ironwood is the new Google hardware.

Gary Arora

Your phone's always listening though. You got to.

Bill Briggs

You can tell. Yeah. So, real-time nudge on, but either way, the sentiment of it is true.

Gary Arora

It's amazing. So, we are in fourth month of this year. So much has happened already in terms of the updates and the innovation we have seen. We started the year with single agents. Now, we're talking about multi-agents that autonomously talk to each other. We started with LLMs and now we have deep research LLMs that can get you PhD-level insights. So, let's assume we are now sitting in Q4. This is December, we are getting ready for the holidays, and we do a look back. What do you think some of the things that happened this year that you are most excited about in terms of the capabilities it unlocks? Let's start with you.

Mamoun Hirzalla

The most exciting thing to me is the release of the Agent Development Kit from Google because if you look at the offerings from Google today, if you want to build an agent, you can build an agent, few clicks with Agentspace that allows you to go across multiple repositories and in five minutes you're there. You can go also take actions. These actions can be calendar actions or mail actions, or ServiceNow actions. And you can do that, extremely powerful, in very short order of time. If you want to be more sophisticated, you can go and build agent using Agent Builder where you can specify a goal and you specify instructions and you add the tools to it, whether it's an API tool, or a calendar tool, or integration tool, or a data store tool. And you can do that also in 10 minutes, 15 minutes. To get more sophisticated, to get to the level of where true agentic capabilities need to be there, that agent development toolkit is required because it will allow you to be interoperable across different, multiple kind of agent frameworks and then the agent to agent framework, or the protocol, as that gets mature, I should be able to discover, I should be able to inter-operate, and the inter-operation here is not just the integration. It's truly about what is the handshake verb? What do you discover, how do you advertise, how do you do all of these things? And then you may want to negotiate as part of that. These were all the promises that we talked about when CORBA and web services came along, never materialized. I think there is a great opportunity right now to go and say this potentially could materialize because there's an LLM behind all of this and these things do really well when it comes to the interaction and the changes. So December timeframe to put the answer quickly, these tools are available now, they will mature very quickly, and I think the right use cases coming out of this will go beyond the hype, the Chatbot, one data source or multiple data sources. You're going to see true agentic use cases that people will be very proud of, and the marketplace will allow you to go advertise. And I hope that we will be able to go and say can I discover, can I use? Without a lot of complexity or sophistication. And that would be probably something that we can get to in December timeframe.

Gary Arora

Very well said. Bill, bring us home

Bill Briggs

Yeah. So, the announcement we made with Google and ServiceNow on the agents that we brought to market today, I hope that we see 100x number of agents that are available with Google and ServiceNow and also, we have partnerships with other tech providers in similar ways. And the power of the continued investment with inter-operability would mean for heart of the business issues for our clients, we're showing up with outcome-based solutions to help them realize it. They're probably going to stitch together a coalition of many different players. Of course, Google is an important piece of that puzzle. But all our clients are having deep relationships with multiple tech providers, and so do we. So how do we weave that to actually evolve and working together for better outcome? And for Deloitte, we'll continue to have amazing people doing great work in professional services, but also more and more investment in our core engineering and platform outcome-based assets. I think agents are going to be at the heart of that. So, we'll continue to see that evolution.

Mamoun Hirzalla

Agree. Since you mentioned ServiceNow, that's close and dear to my heart as part of the partnership. So I'll give you some of the glimpses and the differentiation that could come out of that. Today, when you look at Kubernetes and container workloads and all of these things running Google Cloud, that's one part. ServiceNow has the view of potentially containers along with legacy systems that may not be under the auspices of cloud monitoring as part of Google Cloud. Now, suddenly an issue happens in a container, or things build up to potential issue. An agent is looking, listening and monitoring saying potential issue alert. If that happens, it created a ticket automatically and says I got the issue. We have trained that on a pattern or set of patterns from log data, says here's what the issue is, here's the potential solution. Create, alert, assign and show the solution to it. So, when you fix it, that could take minutes rather than days. The next stage of evolution will be what is the level of trust to see that this level of, exactly the autonomic type of stuff, so you can fix it, and I need your permission to say, can I go and put it into that branch? Go ahead and do it. So that's the vision and some of the action is happening right now, to see this is a big differentiator. This is not just your ordinary way to debug and do these things. This is a huge differentiator, all agentic building on the strengths of Google cloud security operations, the cloud operations, along with ServiceNow to make that vision a reality.

Gary Arora

All right, so now we have you both on record predicting the future. So, let's meet in December and see where we land. Really an exciting time to be in technology and witness all this technology evolution. That's it for our episode. Thank you so much. Bill and Mamoun, you can find us all on LinkedIn and continue the conversation. If you like this episode, be sure to like and leave us a review so we can continue to bring these deep insights from the ground. That's it for now. I'm Gary Arora. I'll see you next time.

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