On Cloud Podcast
September 2023

The Deloitte On Cloud Podcast

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

Title:	**Google Cloud Next 2023: All the highlights, new products, and more**
Description:	Google Cloud Next 2023 was a home run! In this episode, David Linthicum talks with Deloitte's Kashif Rahamatullah about the news from this year's event. AI, and generative AI were the main focus of GCN 2023, with Google introducing more than a dozen AI infrastructure tools to help their customers take full advantage of generative AI. They also discuss how Google's partnership with NVIDIA will enable companies to better leverage large language models for generative AI.
Duration:	**00:12:48**

David Linthicum:
Welcome back to the On Cloud podcast. Today on the show I am joined by Kashif Rahamatullah. He’s a principal at Deloitte with me, and he’s going to recap Google Cloud Next. How are you doing, Kashif?

Kashif Rahamatullah:
I'm doing great, David. It's good to be here with you.

David Linthicum:
Pretend you're a hard-hitting tech journalist and you're going to tell me about what's going on at the event. We're going to have a discussion, get some opinions. Some of this stuff we haven't discussed prior and it's going to be candid and on the go. I did see the keynotes as well. I did so remotely. I've been walking around the show PaLM tree, and I got a different perspective in doing that. What were some of the key items that were announced at the event that you view as setting Google apart in the role in the market right now?

Kashif Rahamatullah:
Yeah, David. I think the theme of the event, as Google has put it, is the new way to cloud, and it was quite evident right off the bat with the keynote that Thomas Kurian and Sundar Pichai participated in. It's almost, in my mind, a pivotal moment. I mean we've been all hearing about the hype, if you will, around generative AI, but I think what we saw in action, because there were quite a few demos during the keynote that you probably noticed, was putting all of that generative AI into action.

Google has been working on and sort of almost created, invented, the transformers, and now you see all of that in action through the power that they're delivering through Vertex AI and the announcement of Duet AI for both workspace and soon to be coming in a GA around cloud. So, the application of those two solutions, the fact that Vertex AI will now incorporate not only Google's own LLMs, but also the third-party large language models, were some key announcements.

I think there was a lot of excitement post that keynote as I kind of noticed the cloud was almost secondary, and gen AI powered by cloud was the key theme of the conference so far, the big announcement with NVIDIA around the Google partnership, leveraging NVIDIA GPUs and the H100, and how that's going to power some of that generative AI capabilities for Google.

Their engineers working directly side-by-side with Google engineers was I think also a pretty significant moment as part of this conference. Lastly, I will say that the Responsible AI team, the fact that Sundar mentioned that in his portion of the keynote, and then it was demonstrated through some actual implementation of digital watermarking on images that are generated through generative AI, that you cannot see and it doesn't distort the quality of the image, was another I think pretty important moment in the actual application of responsible and reliable and trustworthy AI. Those were some of the key highlights that I noticed, David.

David Linthicum:
I think you're spot on. I also noticed the fact that cloud was really not mentioned as much in the keynotes, even though it's a cloud conference. I think we're kind of entering the age where what runs on the cloud is more important than the concept of cloud. We know we got storage right and compute right, databases, things like that. We've been at that for about 15 years and it seems to be working, so that becomes foundational, where these incredibly innovative technologies sit on top of that technology.

I think you kind of hit the nail on the head, where the cloud is becoming a platform, a core platform for AI, and that's really going to be the focus of a lot of the cloud conferences moving forward. I've been to a couple in person over the last months. Really, everybody is pivoting, not necessarily away from cloud, but using cloud as foundational technology. So, on the NVIDIA stuff, what would be a good application for that? What would be a good use case for a typical enterprise to leverage that technology?

Kashif Rahamatullah:
I think as enterprises go through this decisioning processing of do we create our own large language models, do we leverage third-party, first-party large language models, I think it's a way of powering that decision. Once you've made that decision, NVIDIA's GPUs really provide the computing power that you require. Google is of the opinion that you should really first take a serious look at the third-party, including their own LLMs, as well as hundreds of large language models. The other question is do we actually use a large language model? Do you start with a smaller version? Google has now generally available Med-PaLM, which is a medical version of the PaLM large language model.

Similarly, they have a Sec-PaLM, which is focused specifically around security and security threats. They are talking about creating a lot of these specific industry, sector-specific large language models that would allow companies to take advantage. Instead of having large models that require a lot of compute power and a lot of storage and a lot of time to tune, start small. Leverage your own corpus of data to tune that. That's been a key differentiator for Google, because they don't require those models to feed data back outside of the four walls of an organization. So, again, from a reliability, from a security, privacy, and trustworthiness perspective, the model, once it's used, it's tuned, it stays within the four walls of an organization. In some cases that's the power of GPUs and TPUs will come into play.

The other thing is just the engineering partnership between NVIDIA and Google, where their engineers are working side-by-side to tune the models to run effectively, in an optimized fashion, is the other benefit that customers will actually gain from this partnership, because these models will be tuned to run in an effective way. There's a lot of compute power required. So, how do you optimize these models, so that your compute bills are not outrageous? So, I think there's a lot of goodness that will come out of this collaboration and alignment between Google and NVIDIA.,

David Linthicum:
I think the important aspect to that is the ability to componentize these LLMs and use them in certain circumstances. I do see you have mentioned a few already--verticalized LLMs, where you get some general understanding in terms of a verticalized knowledge base, say around manufacturing and health care, but also the ability to build your own LLM and your knowledge base within your four walls that are going to be proprietary to you. If you think about it, this tiered AI is kind of where everything is going, so it's not going to be leveraging a single LLM. It's going to be leveraging multiple LLMs, some that are

owned and provide general purpose information that's needed for a particular industry and those that are associated with just a business. I think that's the way that enterprises want to consume generative AI. What do you think?

Kashif Rahamatullah:
I think you're spot on. I think just like several years ago, when people started considering cloud, there was a lot of concern around data privacy and data sharing, and do I really put my IP and almost differentiated solutions and applications on somebody else's infrastructure? We're somewhat in that same point around generative AI at this point. There's a lot of conversation about do I create my own large language model, because I have so much data that's residing within the four walls of my organization that I own, that I maintain, that I control, versus—and I think that's really where I see Google has taken a very different path, their focus around creative and enterprise-ready solutions that allows enterprises to be comfortable with touching and leveraging this technology.

Because, again, they have offered for you to take a model, a version of the model, tune it within your own tenants, if you will, of Google Cloud, leveraging your corpus of data, and none of that data ever leaves your environments and your tenants within Google Cloud. I think that enterprise-specific view to address the concerns around privacy, security, control on data has been a bit, in my mind, of a differentiator for Google versus some of the other competitor products. So, I'm fully aligned with what you just said.

David Linthicum:
Absolutely. Let's talk about Duet integrated with Workspace. I saw some really cool demos of that during the keynotes. What should the listeners take away from that?

Kashif Rahamatullah:
I think Duet from a Workspace standpoint, just in terms of workforce productivity, will certainly be a significant uplift for Google and Workspace both. But I think what you also saw was how it's going to be now also integrated with cloud and what you can do with it around code quality, code generation, just the speed with which you can generate code. Again, going back to sort of the natural language models or the natural language processing, if you will, so leveraging just plain old English to generate pieces of code that can get you accelerated by days if not weeks is going to be super-powerful. You're also, in my mind, shifting the role, if you will, of a developer and/or an architect and an engineer, and really opening that up to people that may not be as deeply technical in their ability to spin up applications and the use of technology.

So, I think this is stuff that we've been in the space of AI and the space of code generation and code quality reviews and stuff like that that we've been talking about, but now it's in action and it's really awesome to see that. I also want to call out what we saw around Alloy DB, which was another announcement that Google made, and just the opportunity there to modernize and really migrate your data and database applications over to cloud and the speed with which you will be able to do that, and how between Duet and Alloy you are able to leverage the tools to accelerate that process significantly.

David Linthicum:
Yeah. I was kind of taken aback by the fact that we can finally develop applications at the speed of need. We can think it and get to an application state in a very short period of time, and get to a state where we're not just building something that's a prototype that we throw away, but something that's workable that we're able to get into production in a record amount of time to start adding value to the business. I think that's been missing from a lot of the way in which we do development and deployment. Even with the advent of DevOps and Agile technology, it still takes a while to build these systems. Now, leveraging AI is a true force multiplier. We're able to get them out as quickly as we can. So, where can people find out more about you on the web?

Kashif Rahamatullah:
Certainly I am on LinkedIn, and @Kashifvc as well on Twitter. Our relationship with Google has been a multiyear journey with us. We, by the way, were announced as a Partner of the Year as well, Deloitte was this year, so Deloitte.com.

David Linthicum:
Absolutely check it out, a good alliance, a good relationship, a good partner, and bringing lots of solutions to bear for people that need this technology, which I think is going to be most enterprises out there. 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 by 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. So, until next time, best of luck in your cloud journey. You guys stay safe. Cheers.

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 © 2023 Deloitte Development LLC. All rights reserved.