The Deloitte On Cloud Podcast

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

Title: **Deloitte's Tony Witherspoon on all that's new at re:Invent 2023**

Description: In this episode David Linthicum chats with Deloitte's Tony Witherspoon about all the hot topics at AWS re:Invent 2023. They agree that Generative AI was the star of the show this year with AWS announcing Amazon Q, CloudWatch Application Signals, and AWS My Applications—all of which help organizations leverage Generative AI more effectively over diverse cloud architectures. Dave and Tony also discuss the future of Generative AI and concur that 2024 will be about taking Generative AI to the next level.

Duration: **00:23:27**

David Linthicum:
Welcome back to the On Cloud podcast. Today I am joined by Tony Witherspoon, CTO AWS Alliance, and today we're going to recap and discuss this year's re:Invent How you doing, Tony?

Tony Witherspoon:
I'm doing well. As well as you can be from coming back from Las Vegas on the red eye but made it back home safely, so all is good.

David Linthicum:
Yeah, those are tough. Hopefully you can take a nap after this, but I appreciate you coming on the podcast and giving us a recap because a lot of people go to re:Invent every year, it's kind of a pilgrimage, as I call it cloud computing Woodstock, but a majority of people in the industry can't make it, and so they love to get recaps and lists of what actually occurred because it does define a lot that's going on in the cloud computing industry in general. And then I think that kind of the evolution that occurred over the last ten years is we went to vendor-led technology strategies, so we're following vendors, in essence.

So the innovations that they do, enterprises have a tendency to align behind them, so it's an interesting kind of switch that occurred in the last 20 years, and so in the large vendor conferences, which used to be just things that were in hotels and the other things were done by standards, consortiums, and magazines and things like that. Now it's back and the vendors are the primary make-or-break events going forward, and I suspect this was no different. So, what was the energy like there? What was the feeling? There's always a feeling there that's always different from the year before.

Tony Witherspoon:
Yeah, so coming off what we call the post-COVID, after COVID was, like 2021, it was a smaller conference. Last year it got bigger. I think I've heard numbers around 50,000 attendees last year. And then when I was talking to our partners over at AWS, their estimate was probably about 70,000 attendees this year. And I totally agree with you that I feel like some of these re:Invents, they had different themes, like if you kind of look back, it was all about when they initially started providing infrastructure as service, and then there was one year where it felt like the theme was all about serverless when they introduced Lambda, and then it was IOT and things like that.

And definitely this year and probably at most technical conferences, it was around AI and Gen AI and large language models and things of that nature. That was the feeling around the conference. And I guess the other thing I noticed was I met a lot of people this year that this was actually their first re:Invent, or maybe their second one, so a lot of new entries into the AWS marketplace.

David Linthicum:

Yeah, I know a lot of people are planning for the next re:Invent in 2024 this week and getting back because I think if you're in the industry, it's kind of like if you're going to attend one event, if someone's going, “What's the one event to attend?” I would tell them re:Invent’s probably the smartest one to attend because everybody's going to be there, I understand it's led by a vendor, but there's a lot of free thinking going on in how to leverage the technology.

And kind of my feeling this year was that it was a little bit more open, so it didn't feel like a walled garden as much anymore. It seems to be reaching out to partners, even though they're cloud providers, it seemed like they had some multi-cloud offerings that are starting to move up. I did see that the cloud management platform that they use is now going to extend to control other clouds, which is kind of a game changer. That just kind of fell off everybody's radar.I didn't see anybody retweet that or anything else, but it's kind of huge. In other words, we're actually managing a heterogeneous multi-cloud environment and can do so with some tools that are being produced by a cloud provider. What are your thoughts on that?

Tony Witherspoon:
I think that's definitely not just AWS but other partners, just you have a multitude of different customers and different customers objectives and needs, and it's all really about meeting them where they are. So, be it multi-cloud, I had a lot of discussions around AWS has the virtual compute and things like that, but then sometimes customers really need to get down to the core infrastructure, the compute and GPU layer, and so had a lot of customers talk to me about, it was launched several years ago, but AWS's bare-metal instances and things like that, just there's so much analytics and computing power that you need, and to try to get as close to that as possible. So, AWS comes out with these services every year, and so it's not just kind of like this one-size-fits-all. It's basically fit for purpose, and so if you have a need, they have a service, or if they don't have it, they'll build it for you.

David Linthicum:
Yeah, and I think that's a step in the right direction. I think every year I would do a prediction what was gonna be announced at re:Invent, and it was always like I think they're going to open up their infrastructure and become more open and inter-relatable and more open integration with other systems than they are now and kind of move away from kind of the walled garden view in the fact that we have to exist in ecosystem. Obviously we need to play around with existing legacy systems and even work with different cloud providers, and now that we have Generative AI kind of moving in there, it's going to become even more heterogeneous with lots of things that reside on premise and maybe the data stays on premise and the LLMs stay in the cloud.

And, so, we're getting into kind of ubiquitous computing world where any platform is going to be fair game to run your applications, and we're going to optimize the platforms that we run based on the solutions and applications. Sometimes that's going to be AWS; sometimes it's going to be something else, and edge computing, mobile computing, things like that. So, we're moving into this very complex world, and I think that organizations or technologies—AWS probably the primary provider there—who understand that and open up their infrastructure to make that happen and enable their customers to be successful are going to be successful unto themselves.

So, one of the things that also was a pleasing thing even though Generative AI was kind of the star of the show, it wasn't really like the other cloud conferences I was monitoring remotely, a Generative AI conference. It seems like all cloud conferences really just kind of focus on Generative AI, which is fine. I understand why we're doing that, but you've got to remember that most of the rank-and-file enterprises out there are interested in next generation storage capabilities and compute capabilities and database capabilities, and AWS seemed to put enough focus on that where those people didn't feel like they were left behind like they were at some of the other conferences. What are your thoughts on that?

Tony Witherspoon:
Exactly. And, so, one of the things that I really enjoyed at re:Invent is Werner's keynote, and he definitely talked about Generative AI a little bit, but then he did pay homage to just AI in general. And, so, he kind of went back and said, okay, Generative AI is cool, large language models, but there's all these other capabilities that we were doing with machine learning and AI that we shouldn't miss the boat on. So, when Generative AI came out and it was all the buzz and it was like, okay, how can we use large language models to do X, Y, and Z, but that might not be fit for purpose, and so he gave out a number of different examples like drones and things like that that doesn't necessarily need Generative AI, but it needs AI.

And then I also remember kind of talking with colleagues, it's like how do we define AI, like what is it? Isn't it just software and algorithms and things like that what makes AI? Because a lot of times we do a lot of brainstorming with—internally or with clients—and we come up with AI use cases, and sometimes they're just good business requirements or goals and things like that, but they're not—they don't necessarily need AI, and so that's kind of like what we have to kind of balance when we have these new technologies and there's a lot of hype around it is not try to go find solutions or find problems for this technology to fit. Let's go back, look at the business, and try to figure out what technologies we can use to help progress the business.

David Linthicum:
Yeah, and I think that's the big risk moving forward with Generative AI. I kind of went through the AI revolution, which was my first job out of college back in the '80s, believe it or not, and it was very hot for a few years and just kind of went away. And what happened was people were trying to overapply it using different business cases. AI had no value in being applied. And we look at a lot of the applications out there that you certainly can enhance them by leveraging Generative AI as a foundational technology for the application solution, but it doesn't fit everywhere.

And I think that in many instances—and I see this going on now, people are kind of force-fitting what they think is going to be the cool technology to any number of business cases, even if it's just doing transactional sales recording and things like that. AI adds cost. It's going to cost more—the compute, some of the specialized processors that are starting to be released now, if you're looking to use those, those are going to cost more. And certainly the training data and the storage data is going to cost more, and even cost more to run an operation. So, we kind of have to pick our battles here, don't you think? Was that kind of well understood at the conference or did people say AI all the way, we're going to apply it everywhere we can?

Tony Witherspoon:
I don't think we're there where we've got to pick our battles. I think it's definitely still AI all the way. That's where all the excitement is, I think that's where a lot of company investments are. It's probably a whole lot easier to get investment approved if you put the word AI somewhere in the abstract. So, I think that's where it is, and then I think the overused analogy of where I think we should be thinking about is kind of scanning where the puck is going to be, not

where it is today, and I think that's where other enterprises and technology companies like Deloitte, we can help with that because I think a lot of the problems that we're solving with Gen AI are either probably not the right use cases or the easier ones, but the where we need to be in three to five years is where I think we need to spend a lot of our focus on.

David Linthicum:
Yeah, so the value of a trusted advisor is one who's technology agnostic, doesn't have a dog in the hunt, so to speak, and the ability to kind of pick the right space and look at this objectively. I think ultimately those are the companies who do that are going to win the game. That doesn't mean they're not going to leverage Generative AI, but they're going to target it at a certain number of problem domains and certain level of use cases, and I think that's really key to winning all this because you can't do everything.

And also, there's enterprises out there that have absolutely no plans to move to Generative AI systems, and so they're just trying to keep their transactional sales-order systems, manufacturing systems running. I'm sure some of them are walking around the conference as well, and you've got to kind of make sure that we don't leave those behind because that's kind of core to what makes cloud computing valuable, the ability to replace existing on premise infrastructure with something that's going to be held on demand that's going to do a better job and is going to be the best, more optimized path to make it happen.

So, let's talk about the specific announcements. Obviously, we're not going to go through all of them. You can read that online. What were some of the ones that excited you?

Tony Witherspoon:
I'll tell you some that excited me, but when I was having a discussion at the conference and we talked earlier in the discussion about themes, this one the theme was definitely AI, however, it felt like when you watch one of those movie trailers, and you watch a trailer, you felt like you watched the whole movie but just watching the trailer. I already knew or kind of assumed that the theme was going to be AI, and so that part didn't excite me as much. But one of the things that I really liked is we do a lot of previews with AWS and different services and things like that, and so seeing a preview prior to re:Invent and then seeing the announcements and seeing how it's evolved even in that short period of time is pretty exciting.

So, one of the things we had—I think a lot of people talked about was Amazon Q and so how it's kind of built into the console, how it can support in a number of different areas like a built-in IDE and things of that nature. And Werner also talked about it as well, and I had this feeling just kind of working with AWS over the last dozen years is that when I first started working with it, it was probably like 12 services, and I can't even tell you how many number of services now, and he just said on stage, he's like, "You can't be an expert in all these services, but Amazon Q can be. It can be your sidecar along with you, and you can ask it questions so you don't have to be."

You can be broadly across all these AWS services, but if you need to get deep, you have this sidekick, Amazon Q, to help you out with that. And, so, lots of people were talking about it. looking forward to over the next couple weeks as we get to the holidays to kind of dig deeper on some of these announcements.

David Linthicum:
Yeah, that's going to really kind of define how we interact with these systems, the ability to communicate, and that's something that I think the other players are getting right, and I think AWS has to get that right. They understood that, so that's why we saw Q release, but it's going to be interfaced into how we do collaboration, how we do coding, and it's going to be a number of ways that we get in and access the system. And I like the fact that instead of talking about kind of the core co-capabilities, they talk about the underpinning infrastructure as to what Q is and how it can be applied in so many domains and the ability to kind of take your business to the next level by using this particular technology. And, so, I thought it was great for them to focus on that, and I think that's going to really kind of enhance their use of Generative AI, and other technologies for that matter. So, what were some of the other announcements?

Tony Witherspoon:
I also like when I was talking with Adam and he had the talk with the CEO from NVIDIA, and it kind of gives us balance because when AWS first came on scene, it was all about the infrastructure services and things like that, but as things evolved, it got closer to the business and business capabilities, and so how to accelerate people. So, Amazon has all these building blocks, but then also Amazon builds services leveraging those building blocks. And, so, when they talked about all the other announcements, like with H200 chips and how you had to interlink between the different GPUs because if you need multiple GPUs in a single instance and things of that nature, and so it kind of links back to, okay, homage or not homage but how Amazon has built these core infrastructure systems, it balances that with then also straight to consumer type services like Amazon Q where you can use that to leverage insights across your internal enterprise. So, I really love the beginning of Adam Selipsky's keynote with NVIDIA because it kind of brings you back down to the roots of why AWS exists today.

David Linthicum:
Yeah, and also the fact that we're building things out of component parts and they're talking about the component parts not necessarily the cool interface layers, things like that, which I thought was probably overemphasized in the other cloud conferences this year. And the great thing about Amazon is they really kind of focus on how you build and deploy this stuff. They're obviously a company of developers and makers, and that kind of comes out in how they look and present their products. And, so, the piece part component stuff from the very sophisticated down to the lower primitive levels, how you leverage these building blocks in certain ways to create unique bespoke solutions that are going to have value specifically to you, not something that's going to be all in and things like that, I just think it's a better approach to doing it, and that probably gets down to their success. So, what other announcements kind of got you excited?

Tony Witherspoon:

So, not necessarily an announcement, but Werner did this one thing where he took this moment to say, “I keep telling you that using these building blocks are easy, and you can use our services to enable your business quickly,” and so he went through this story where he went to go find a problem, and going back to his roots of working in radiology and what radiologists are—what the main issues, and so you had a bunch of conversations with them, and he built a model within SageMaker to scan images and to find out if someone's had a stroke as early as possible because every second that someone had a stroke, they're losing brain cells. And, so, he kind of walked through it and showed how he started from beginning, imprinting images into SageMaker analyzing it, and producing these results.

And basically saying, “If I can do it, then you guys can do it.” And, so, I thought that was pretty cool because it kind of takes a personal story into where are you at, and I'm not just saying these are easy services because I work at AWS. I've actually gone through that process. And, so, it kind of speaks to me because as I progress through my career, I get to spend less and less time on hands-on with the technology, which is kind of why a lot of us got into the technology business because we like solving problems, we like coding, we like trying things out, we like seeing—not necessarily seeing them fail, but we like when we see those error messages, actually fix them. And, so, to me, just kind of speaks to how easy it is to use these services, and then also it's good to go back and kind of experience those things that you did at the beginning of your career as well.

David Linthicum:
Yeah, and I like the theme The frugal architect. I think that's a great message to send because I think we're at a point right now, and we made a lot of mistakes over the last four years certainly around the pandemic where we may have moved too quickly to the cloud and not done the things to make our systems as optimized as possible, and I think I look at those things when they saw the big bills that came in 2022 , and most of the time those are self-inflicted wounds because we didn't have a frugal architecture kind of mentality when we went at this, and it's kind of refreshing you see someone who's a CTO in a public cloud company to kind of stress the fact that at the end of the day we need to get down to the fundamentals and how we build and deploy these things.

So, even though the technology layers on top and certainly Generative AI stuff is cool and neat and all the cool kids are going to do it, I think we're moving back to—and I look at this in 2024, optimization trend where we're going to not only build net new systems that are going to be more highly optimized, and certainly we have some tools, even the Gen AI tool stuff allow us to hone and measure these things, but the ability to go back and optimize existing systems. I mean, we call it modernization, the ability to kind of take everything into a cloud native infrastructure and things like that, but we built up so much technical debt in many instances in moving to the cloud, that seems to be the next path moving forward.

That doesn't mean we're not going to continue to move to the cloud, but we're going to go back, loop back, and optimize these systems that are under-optimized, and certainly when we move forward with Generative AI, they've got to be optimized because they're going to just burn too many resources if they don't, so we need to think about how we do things in a frugal, sustainable, growable way, and it was key to me that that message is coming out of an AWS conference.

Tony Witherspoon:
Yeah, so there were a couple announcements that—new services that haven't had a chance to look through that that may address some of those. So, there was CloudWatch Application Signals, and then also AWS My Applications where it kind of does a lot of this observability for you, so you don't have to do it from scratch. And, so, if that's anything that anybody's interested in, I would definitely go look into those services. And then someone kind of jokingly, we had a chat going between—because some people were actually at the venue watching keynotes, some people were at different watch parties, or they might be sitting in their hotel room streaming it. When Amazon Q came out and it's integrated in a console, like wouldn't it be good if I can just, in natural language, ask Amazon Q how to optimize My Applications, or what's the cost and what are the key things and create dashboards and things like that where you don't have to be an expert in cost explorer or use expensive third-party tools to do that and it's just built into the application using AI.

David Linthicum:
Yeah, that would be amazing. I think that we, like I said, we're doing self-inflicted wounds maybe because we're moving too fast. The ability to do those things fast and come to those conclusions quickly, and even the ability to integrate with things like FinOps or forward we're going back and looking and second guessing a lot of the architectural changes that are made and even looking out to the different technologies and having models that are able to understand what's going on currently and the ability to look at the opportunities to integrate that technology in your existing applications and solutions. I mean, it's going to make—I'm an architect by trade. It's going to make our jobs easier moving forward, and it's going to put—it's going to federate a lot of stuff that needs to be federated in the hands of the people that really kind of make a difference. Anyway, final thoughts on 2023?

Tony Witherspoon:
Yeah, I think it was AI, AI, AI, and then 2024 is all about how do we leverage this in the best way possible.

David Linthicum:
Yeah, and I was going to ask you what do you think we're going to be talking about next year at this time when we do the recap? What do you think—putting you on the spot here. Is it just going to be using these technologies effectively or is it going to be some kind of a net new technology à la Generative AI that we're going to be talking about next year?

Tony Witherspoon:
Honestly, I don't think it's going to change. I think we're going to be advanced a little bit more on this AI side. I guess the other thing—and it's been over the last couple years we've been talking about industry led solutions, so how do we use AI for automotive or retail, consumer products. I think it's going to get more focused in that area as we—or financial service, like how do we help our clients directly leveraging these technologies.

What I've seen in Deloitte, we've done a lot of proof of concepts and prototypes and things like that. Hopefully, by the end of 2024 we'll actually have more tangible use cases of how we leverage these technologies. And then I also think if we can leverage it in one industry, other industries are different, but we can kind of take a 20 percent pivot and kind of use some of the same patterns or insights and help other areas as well.

David Linthicum:
Yeah, I think that's a sound prediction. I think 2024 is about us getting this right and the ability to have pragmatic use cases for all of this technology. I think some of them are being implemented now, but the sky's the limit in terms of supply chain optimization, the ability to apply these particular vertical spaces, finance, healthcare, things like that, like the healthcare use case that Werner's had, but I think the sky's the limit. And, so, I think we're getting to a healthy state of cloud computing. Around 15 years, people are starting to use it, they're starting to make mistakes, but we're starting to course correct and starting to layer technology on there really is going to make a difference. So, where can we find out more about you on the web, LinkedIn, profile, social media that you do?

Tony Witherspoon:
I don't do a lot of social media. I am on LinkedIn, so I try to communicate on that as much as I can.

David Linthicum:
Yeah, reach out to Tony. He certainly knows the AWS infrastructure, but cloud in general, so make sure to follow him as well. So, if you enjoyed this podcast, make sure to like us, rate, 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. If you want to contact me directly, you can email me at dlinthicum@deloitte.com. So, until next time, best of luck with your cloud journey. 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.

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