On Cloud Podcast
November 2023

The Deloitte On Cloud Podcast

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

Title: **David Linthicum on how Generative AI fundamentally changes cloud architecture**

Description: In this episode, David Linthicum discusses how Generative AI changes cloud architecture and what organizations can do to leverage Generative AI effectively. As David explains, the fundamental goal with Generative AI is to return as much value back to the business as possible with the implementation. This search for value requires re-evaluating the cloud architecture from top to bottom, from goals and objectives, to security, model selection, and scalability—and everything in between.

Duration: **00:18:45**

David Linthicum:
Welcome to this Deloitte On Cloud Podcast Knowledge Short, exploring a specific topic related to cloud computing. This is a short tutorial talking about real-world concepts in the emerging world of cloud computing. I'm your host, David Linthicum, cloud computing subject matter expert, author, speaker, and managing director with Deloitte Consulting. This is, "How Generative AI changes cloud architecture."

This is a common question I've been getting for the last six months. Everybody is cycling over to building cloud-based systems that are leveraging Generative AI capabilities. Of course, the question would be: How does it change our architecture? What are some of the changes to the processes that we normally would do when building a cloud computing architecture? And, really, kind of just generally what's the difference? What changes? What kind of skills do we need? All these sorts of things are pertinent. So, I figured I'd do a quick tutorial and walk you through at least everything I know about it, and maybe you'll learn something, as well.

So, first and foremost, you need to understand your use cases. I think the biggest mistakes that people are making with Generative AI systems is that they're over-applying the technology. I understand it's very cool. It has great capabilities. It's fun to work with. Artificial intelligence systems have always been fun to work with. By the way, they're not new. They have been around—I was working on them when I was 18 years old, and I'm not 18 years old anymore. However, the capabilities and the cost of running them is what has changed. So, the low-hanging fruit seems to be: How do we leverage this technology in a particular application? And, in some instances, it's going to be applied with business applications, really, where it really shouldn't be applied.

At the end of the day, Generative AI systems are going to require more processing and more storage, and therefore they're going to cost more. So, there has to be a valid business reason for leveraging this technology. Kind of keep that first and foremost, because I suspect that many of the organizations out there that are going through transformation of existing applications to leveraging Generative AI, or even building net new applications that are leveraging Generative AI, are probably doing so without a business case and understanding what value is going to be returned back to the business.

So, you need to have a clearly defined purpose and goals of the Generative AI system within your cloud architecture. In other words, everybody understands why we're leveraging it and it's easy to explain the business value that it's able to generate. So, I would write down and find consensus on the objectives, address the goals and define success. So, with the larger team and with your leadership, everybody understands why we're doing this and why we're leveraging this technology, and it's clearly defined in terms of the business value that's going to be returned back to the business.

There's a huge amount of value that you can build with these systems, because if we're able to provide better customer experiences through the use of Generative AI systems, if we're able to provide a better product or service through the use of Generative AI systems, that goes directly to the value of the company as well as the bottom line of the revenue. So, lots of benefits to be made, but you have to pick your battles.

Next, keep in mind that data sources and quality of the data are going to be key. Identify your data sources required for training and inference engine by the Generative AI models. In other words, what data is going to be needed to train these systems? They're only as good as the information that's training them, and if you're finding that the data has lots of erroneous information, it's not good quality, then you're not going to get good responses and good knowledge that comes out of these Generative AI systems. These knowledge models aren't going to be built with the correct and proper data, and therefore they're not going to able to add the value. Data is key for all of this. So, keep in mind that data accessibility and data availability is going to be key to making this architecture work, and it has to be first and foremost in your cloud computing architecture.

Next, focus on data accessibility as a primary driver for cloud architecture. We just mentioned consider efficient data pipelines. How are we going to get at the information that we need, where it exists, and build the knowledge models on top of that data? In many instances, I think that the temptation is to migrate everything to a centralized data repository, such as pushing everything onto a public cloud. I think in some instances that's going to make sense, but in many instances, we're trying to leverage the data where it exists. So, if we have transactional data, business data, sales data for example, inventory data, things that we're trying to leverage to train our knowledge models, then just the ability to get access to the data is going to be good enough.

We don't need typically to relocate it or create some sort of data warehouse or auto-migration system that occurs. That's going to be very expensive. Ingress and egress processes are pushing costs. They're pushing data into the cloud and pulling data out of the cloud. It's going to be very expensive. In many instances, it should be avoided as much as you possibly can because it's going to make cloud computing much less of a value than it actually is. So, that needs to be considered. The data is everything with this. So, data sources and the quality of the data are going to be key to a successful Generative AI system in the cloud.

Next would be data security and privacy. You need to understand that you need to have robust data security measures such as encryption, access control, multi-factor authentication, things like that, and it's even more important when leveraging Generative AI systems. Not only do we have access to very sensitive information that may train these LLMs, large language models, but you're going to have to make sure you're limiting access to not only the training data, the data that's used to be consumed in the Generative AI systems to train the systems, but also the outcome data.

Keep in mind that Generative AI systems can create very sensitive information, very anecdotal information. In other words, you may have lots of raw data that unto itself probably is useless. It doesn't really mean a lot to anybody that's going to look at the data. However, if it's run through a Generative AI system, we can create summaries of it. We can find information in that data that may in turn be sensitive. So, understand that you may have non-sensitive data that goes into these systems, but sensitive data that's generated and needs to be protected using appropriate mechanisms.

Make sure you comply with relevant data privacy regulations and architect security into the system. So, basically, we have to build the system from the ground up to provide the security that it needs. So, this is not going to be another instance where we build a cloud computing architecture and then we try to bolt a security system on, on the back end before it's deployed. It's going to have to be engineered into every step of the development of the Generative AI-based systems. By the way, that's no matter if it's on a public cloud provider or not.

Next, scalability and inference resources. Plan for scalability of cloud resources to accommodate varying workloads and data processing demands. This is going to be about consuming more processing power, certainly GPUs and CPUs, which are kind of core to the engine of what Generative AI does, and also information or data that needs to be stored, either data that's consumed in through the Generative AI system or is generated by the Generative AI system.

So, you need to understand where your processing is looking to go. You need to understand how to tune and adjust that processing so it's as efficient as it possibly can be, based on the profile of the Generative AI system you're going to be running in the cloud. So, we can't just open up all resources to be consumed by this system, because that's going to be way too expensive. So, we have to figure it out in a mathematical form, and you can do this, what the amount of GPU utilization is going to be, TPU utilization is going to be, storage utilization is going to be, what kinds of processes you're going to run on, the amount of memory you're going to use. All these sorts of things should be built into the process of building your Generative AI system that's going to run in the cloud, because without understanding that, we're going to deploy lots of systems that are going to generate huge cloud computing bills, and that in essence is going to remove a lot of the value that comes out of building and deploying these systems.

Next, model selection. Choose an example of Generative AI architecture based on the use case and requirements. So, we're building something using a model that we're going to build, knowledge models, LLMs, and in many instances we're going to use existing LLMs that other people are providing. For example, if we're in a particular vertical market space such as finance, we may leverage a large language model from a finance vertical to understand or have access to key information, key knowledge bases around that particular vertical, such as how we do risk analytics, how we do compliance, all those sorts of things and those are going to be extremely helpful.

So, part of this, and this kind of gets down to the Generative AI-specific capabilities, is to look at the use cases and look at the model selections that you need to leverage to get to the objectives that you've already set. In other words, the use cases really define the usage of the system and where you're going to get to using this technology. Really, that's Generative AI engineering that has to be done.

So, you need to implement a robust model, deployment strategy, versioning, and containerization in many instances. So, robust means that we're going to put something together that is able to optimize the use of resources, but it's not going to overuse resources. It's not going over-provision resources. It's not going to use more resources than it actually needs. Or we're not going to under-resource it. In other words, it's not going to starve from processing issues and storage issues because we're not providing the resource that it needs. It has to be just right. In many instances it has to provide scalability that aligns to the particular needs and changing needs of the Generative AI system running in the cloud.

Next, monitoring and logging. We have to operate this system, so we have to set up operational capabilities as we build and engineer this system, very much like security. So, we need to set up monitoring and logging systems to track the AI model performance and resource utilization. So, it's one thing to set something up and hopefully engineer it correctly, so it's as optimized as it possibly can be, but it's another thing to look at it as it runs, to make sure that it's living up to the utilization and optimization expectations that we set for it, and we're able to adjust it as it moves.

So, we have to monitor. We have to log these systems. This is when AIOps, artificial intelligence operations, which I did a Knowledge Short on a while ago—check out that podcast if you want to understand more about AIOps—but the utilization needs to match directly the needs of the application as much as possible. So, leveraging things like AIOps allows us to leverage the concept of observability, which means we understand not only what the operations is doing, but we're not looking at massive amounts of noise or data, raw data that's coming off of these monitoring and management systems, but we have layers of capabilities able to tell us what the data means. In other words, it will tell us that we're overutilizing a particular GPU cluster. It could tell us that we're about to run out of data space and we need to provision more.

All these sorts of things are very helpful to those who are charged with operating a Generative AI system in the cloud, because it's going to be a very challenging thing to operate on a day-to-day basis. We're going to have automated tools and technology to make this much easier, but still, the human beings have to figure out what needs to occur in some of the larger decisions that have to be made. So, establishing a learning mechanism for anomalies and use cloud cost management, good FinOps tools, to make sure we're aware when things are going wrong, and we're also able to self-heal and, in essence, correct them automatically, correct issues automatically such as providing more storage if we're about to run out, and it's the middle of the night and no one is going to be able to wake up and make that happen.

Also, cost management tools, the ability to have a layer of FinOps technology that's able to monitor the costs of this technology as it's running in the cloud, and make sure we're utilizing it correctly and we have accountability, and we have observable of the cost systems, very much like the operational observability, where you figure out where the costs are coming from and what's actually happened. All those things are very important.

Keep in mind that a FinOps system is important to a larger cloud deployment overall, and you really should keep that in mind as you build these things. So, FinOps is going to be a cross-platform system that's going to operate across cloud providers and has operated cost and Generative AI systems. This is just an instance of a system we're building, but we need to layer that FinOps system into this particular application instance, in this case, an application that is using Generative AI capabilities.

Some of the other considerations would be you need to ensure high availability and implement BCDR, business continuity and disaster recovery plans for failover and redundancy. In many instances that may go without saying, but I see a lot of systems out there that are built without this in mind. In other words, they really get dependent on their AI system. Suddenly there's an outage, something goes wrong, a cloud outage, a network outage that occurs at a particular location that happens to be producing training data. That training data is not available, and we have to figure out some way around leveraging the system in such a way of bringing the system back up and running, where we're able to get the value of that system ongoing.

In many instances, these are going to be the systems that are going to cost a million dollars an hour in terms of their unavailability. In other words, when they go down—banking systems are an example of this—we're losing massive amounts of dollars. So, we need to kind of adjust the fact that we're going to have to build these capabilities, these BCDR capabilities around the functions of the system and the purposes of the system and the importance of the system, and make sure we're not overspending for creating redundant systems, but we're spending up to what's needed in terms of how much damage is going to be done if a system outage occurs. It's not an easy thing, but it's very important I think that the architects do that.

So, what about cloud computing architecture in general? Some aspects are the same when using Generative AI and some are not. So, we're getting to a purpose-built system here that does Generative AI-based patterns of processing. That could be a number of things, including image generation. It could be the ability to do information automation, the ability to automate a supply chain. Just a massive number of applications that are available that Generative AI is able to address.

But we have to understand that that's going to be part of a larger cloud computing architecture. So, many of the disciplines that we talked here, all these aspects and moving components of building a cloud computing architecture. So, we're getting to a point where we're as optimized as possible. In other words, we're leveraging the correct technology in the right configuration that's returning the most value back to the business, and that's the objective. It's not your ability to leverage a certain type of cloud or any number of different technologies; it's about your ability to get to a point where we're ultimately as optimized as we can, and therefore we're returning as much value as we can back to the business.

So, be aware of the importance and rigor required for Generative AI. It requires that you pay more attention, as I found as I'm looking to build these systems and have built these systems over time. I've been working with AI for a long time, and many of the patterns are exactly the same. They can be massive resource utilizers if they're not checked and they're not designed properly. We have a tendency to forget about that because back in the day we built AI systems on owned hardware, so we already paid for the capacity, and we just used the capacity we're already paying for. It doesn't cost more if we're leveraging a server at 80 percent utilization or 10 percent utilization. Therefore, we didn't put a lot of thought into how we were building and scaling these systems.

Now with public cloud being a utility model, in other words we're paying for what we use, massive bills can be generated if we're not able to keep this stuff in check. I suspect that in 2024 and 2025 we're going to hear about a lot of consternation where people are getting these massive cloud bills from their Generative AI systems they built in 2023. In many instances they're self-inflicted wounds. They didn't go through the rigor of what we're talking about here, and doing the right architecture in the right way.

It shouldn't scare anybody. There is emerging information, courses that are out there. College and universities are teaching this. The on-demand training is teaching this. And basically, if you're listening to this podcast, you're getting a jumping off amount of information to go on your learning journey. This is about continuously learning and understanding that the technology is always going to be changing.

Generative AI is certainly important to us building these various systems, and as a cloud architect you're going to have to understand how to make these things work and how to configure and deploy these systems. Or if you're an IT executive or if you're a CIO or even if you're a developer, what's changing? In other words, what do we need to focus on as we're building and deploying these systems? That's really what you should focus on right now, certainly if you have a base understanding of information and you've done architecture in the past. So, you need to figure out what is going to be new.

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 with 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.