The Deloitte On Cloud Podcast

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

Title: **How an AWS financial services specialist thinks gen AI will transform the future**

Description: In this episode, David Linthicum talks with Ruben Falk - capital markets specialist with focus on Data Architecture, Analytics, Machine Learning & AI at AWS - about how generative AI is transforming financial services. They discuss how gen AI will revolutionize processes across the enterprise, as well as how benefits gained will transfer to end customers. Falk cautions, however, that the regulatory environment for AI is growing, and ethical use of the technology is paramount.

Duration: **00:23:17**

David Linthicum:
Welcome back to the On Cloud podcast. Today on the show I'm joined by Ruben Falk, capital markets principal, data platforms, machine learning, and generative AI at AWS. Ruben, welcome to the show.

Ruben Falk:
Thank you.

David Linthicum:
How did you get to that role at AWS?

Ruben Falk:
I started out in capital markets a long time ago. Initially I was an investment banker with UBS. Then I spent 15 years at S&P Global. I ended up heading up the investment management solutions business there, focusing on data and analytics. Then I came to AWS via a short stint on the buy side. So, I've always been focused on capital markets, investment management, data analytics, NLP, and the likes.

David Linthicum:
What do you do during the day at AWS? You work with big cloud hyperscalers. So, how does that fit in?

Ruben Falk:
I'm sort of in the technology business now, which is new for me, but I spend pretty much all of my time connecting business problems to technology solutions. So, I understand the capital markets space, the financial services space. That's where I spent my whole career. I am learning and getting better at the AWS components within that space, and I am now connecting the two. So, I can speak sort of the business language of our customers, and I can then bring to bear the technology solutions and services that AWS can offer.

David Linthicum:
As somebody who has built banking systems and financial services systems many times in my career, obviously AI has been on the outskirts. In other words, we leverage AI for risk analytics and those sorts of things, where it's very specific to some sort of function that's normally high functionality, but high reward in making those systems. So, I guess the financial services area, with the explosion of generative AI and certainly the ability to leverage this at a price that was unheard of just a few years ago, have got to be anxious about getting these systems in place. Am I off base?

Ruben Falk:
Yeah, that's right. I think it's really the first time where I've seen customers come to me and asking me to help them identify their use cases or their problems. Usually, they come with a problem and ask for a solution. Now they come and ask for the problem. That's I think engendered by the fact that generative AI is seen as transformative, that certain industries, the competitive landscape is going to change. Certain moats are going to be broken down or

moved. So, companies are worried that if they don't think about generative AI and the application of generative AI in their business, that their competitors are going to think of it, and they are going to be at a disadvantage.

So, for that reason, the impetus to think about generative AI is coming from the top. It's coming from the board. It's coming from the CEO. And lines of businesses are having to think about what the application might be within their business, so that they can be on the forefront of those developments. So, that's generating a lot of activity that is a little bit unusual in that we're helping with the problem definition as well as the solution.

David Linthicum:
Yeah. I'm seeing the same thing. The edict comes down from the board of directors. They heard that this technology is going to be game-changing. They want to figure out a strategy to make it happen. So, they're the ones pushing for use of this technology. Of course, the people that are in charge of building these operational systems and finding the pragmatic use cases for it are the ones that kind of have to figure it out. So, I love the way you put it. They're coming to you, asking what the problem is. Because use cases seem to be the biggest fallacy, I think, within any kind of an AI application. I think generative AI is no different, the ability to find a reason to use it, where it's going to truly add pragmatic value. It's not just a bolt-on thing that we do as a checkbox to get our generative AI cred, but it actually brings value back to the business and that's a more difficult problem to solve. What do you think?

Ruben Falk:
Yeah, I think that's right. It's obviously not the case that generative AI is the solution to all problems, particularly when it comes to more traditional machine learning problems and solutions that are very numerically based. That's not where generative AI is necessarily strong. If you ask generative AI to add up two and two, you may get the right answer, you may get the wrong answer, but it's certainly not its strong suit. Its strong suit is working with textual data as well as images, although for financial services the focus is really on the large language models and working with textual data.

That's where it's really strong. So, now all this textual data that in the past was sort of less relatively untouched, all of a sudden you can now extract value from, and even extract value from large repositories of text in a way that just wasn't feasible through manual labor in the past. So, to really be able to get value out of all that data that's out there, be it filings, broker research, earnings call transcripts, news and so forth, and be able to access that going back over several years, across multiple geographies and multiple languages, and distilling that down to a salient point that are interested in, that all of a sudden becomes achievable in a way that it really hasn't been in the past.

David Linthicum:
Yeah, it's not only just reading every 10k, but the ability to understand what they mean and what they all mean in context of each other. So, even if you're an avid reader and you're reading a lot of the SEC filings and trying to understand the basics about what a business is and how it's doing, something like this, with the ability to, in essence, analyze massive amounts of information and tell you some of the patterns that it sees coming in, and also looking at historical patterns and how they led to certain outcomes, we can get a better idea as to what this information means and provide a key advantage in the market or just in financial services in general. So, what are the core use cases you see out there?

Ruben Falk:
Some of the use cases are pretty common within financial services. I would even say it's across industry. So, this idea of internal and external customer service by transforming the customer experience through call center technology, for instance, is something that's already been in place for a number of years. You can call into a call center. The call will be transcribed in near real-time, and a repository of documents will be searched based on the queries that the transcript can generate. So, already today, a call center agent can get relevant articles popping up, prompted by the conversation, but they can then only pass on that article really after the call to the customer, maybe read it themselves and summarize after the call. But now with generative AI, all that information becomes available into a summary form in near real-time to the call center operators.

These call center operators now can be sort of super-human to some extent, in that all this information that previously wasn't available to them is available to them, sort of automatically by just prompting at the document repository with the call transcript. So, that's one example. In the call center space, there're lots of others that are in that omni-channel interaction space for banks, insurance companies, and wealth managers. So, the idea that you can summarize previous communications with a client as you enter into a new conversation, so when somebody calls up, you will have a summary of the chatbot interaction there, e-mail interaction to previous calls presented to you as the conversation starts.

So, you can have a much more fruitful, intelligent conversation, because you're not starting from scratch every time. Then also, all these conversations, they throw off data in the form of transcripts, and those can now be interrogated using natural language. So, you can something like, "What were the topics discussed with the customers over the last three months that had the most positive or negative sentiment associated with them?" So, all of a sudden not only can you improve the customer experience, but you can get, much more easily, much more value out of the data that those customer interactions throw off.

David Linthicum:
Yeah. There are lots of pragmatic examples of how we can leverage this technology. I think, though, that they're kind of accelerating in the financial services marketplace, aren't they, versus manufacturing and retail. The ability to find value that traces directly back to the success of the business is going to be around innovation and use of this technology in new ways that probably weren't thought of just a few years ago. So, you create kind of a new aspect to this business. Am I getting too bold?

Ruben Falk:
No. I think that's right. I think certain solutions are cross-industry, but it just happens that, for instance, for knowledge work platforms, this idea that as a knowledge worker you are essentially researching and summarizing text from various documents, be it as an investment analyst you are constructing an investment thesis, as a lawyer, or as a regulatory analyst you are shifting through precedents or regulations, or as a salesperson you are responding to R2s and R5s.

In all those cases, as a knowledge worker you can now be presented with a very good starting point for the ultimate deliverable. You can use generative AI to research to various documents and put together the skeleton and starting point of the thesis. Then the knowledge worker's role changes to that of

somebody checking for completeness and accuracy, and also interacting with the generative AI to add new sections of text and so forth. It just happens that that knowledge worker profile is very, very prevalent in financial services, and for that reason it's a very fruitful area.

Another area that's also somewhat specific to financial services is data extraction from unstructured text. There're so many communications that are very specific to, say, trade confirmation e-mails, or stock lending, transaction e-mails, or term sheets for structured products. These are all use cases where extracting data from that text today is largely manual. Because they're all very narrow niche domains, if you will. It's hard to train a traditional machine learning model on them. You'd have to train one individually on each niche domain, whereas with large language models you can essentially—for the most part at least—you can submit the document or e-mail to the model and ask for the data to be extracted, and it will come back and you're relying on the pre-training of those models.

David Linthicum:
Yeah. The great thing about that is they improve over time. So, you're getting not only answers to summarize your information, but you're getting that information in context of other data points are going to be more useful to you, which is going to be a key success in financial services, and lots of industries for that matter. So, what is the AWS approach to all this?

Ruben Falk:
We essentially offer choice as a first priority, choice and flexibility in the models that our customers can choose from. Obviously, as always, we're focused on data security and governance. By default, we won't train on any customer's data, and we keep all that data private within their own VPN or virtual private cloud as we call it. Then scalability and cost, as a hyperscaler we have that scale. For cost purposes, we also have our own custom silicon that sort of provides an alternative to the standard CPUs out there that are lower cost, and also where we control the supply chain, which is quite important these days with the supply chain issues with CPUs generally. So, that's sort of the basis.

But really, the flexibility and customization are key. So, originally when these LLMs came out and captured the public's imagination, the idea was that you want a really big, really powerful model, and you can ask it any question and you'll get the right answer back. It will pass the bar exam, the SATs and so forth. Now it turns out that a year later, for our financial services customers, that's really not the way they want to use these models. They want to be able to control the fact base on which the answers are constructed. Because if you don't control the fact base, then there's no traceability to the source document, and that means that you cannot be sure that it's even a correct answer or a hallucination.

Without traceability, and if you're in a highly regulated industry such as financial services, you really don't have a marketable product. So, for that reason, the solution pattern or design pattern looks something like this. You have your own repository of documents. You do a search on that repository. Then you feed the search result text snippets into the large language model for additional summarization, paraphrasing, or data extraction. So, with that, you can control the fact base on which the answers are based. Then you can trace lineage back to the source documents. You can check for hallucinations as sort of a second step, because you can ask the question of model again: was the answer provided really based on this specific document? Does it contain the specific entity in the document pattern. That is the design pattern and that makes it a very different proposition than just using a large language model for everything.

Because now, what you have to do is you have to take the large language model technology out of the model and apply it to your own dataset, so that you can search for, as you said earlier, you can search for meaning. You can do a semantic search, so that the model understands what the question is that you're posing and can paraphrase that so that it can search for some relevant answers in your own text repository. So, that becomes the important part of the solution. Then you feed the results to the large language model, but in that case, you don't necessarily need a trillion-parameter large language model for doing something that's very bespoke to your own domain, where a lot of the modeling is happening outside of the model. That's just one example of why fit for purpose models is what we're seeing, and that these very, very large, very, very general models, they have a place, but they're not necessarily always the right answer.

David Linthicum:
What kinds of questions are you getting when people approach AWS and they're looking to leverage generative AI in the financial services area. What are they concerned about? What are they excited about? What do they see as a functional use case for them specifically, not just use cases in general? What are the customers saying?

Ruben Falk:
The customers are excited about the possibilities. So, we have customers that are working on specific use cases in the call center space. It's still early, so we have that many reference customers yet, but little by little we're getting more and more. For instance, principals in financial are working in the call center space. They're talking about it. They're working with our generative AI solutions to construct that more interactive call center experience in terms of summarization for the call center agent.

But they're clearly also having to get comfortable with regulatory requirements, compliance, the privacy of the data they're feeding to the models, and so forth. It happens that we have very good answers and solutions for that, but it's an area where customers are still in search for education. Then there are also questions that are still unanswered, things like copyright. How do you make sure that the model you're using is based on material that's not copyrighted? For a lot of open-source models, that's a difficult question to answer. But again, if it so happens that if you're working in open source and you're working with a fully flexible setup, then you can pull out datasets from the training data that is in fact questionable. So, yeah, still a lot of open questions. Regulations are coming out little by little. The EUAI Act has been proposed and that requires all model vendors to document the source of all the training data, which none of the big model vendors comply with today.

The SEC just put out some guidance around fairness and also a requirement to report when financial services institutions that are servicing retail clients are using what they call predictive analytics, which seems incredibly broad, but I think in part they're certainly targeting AI and generative AI. So, we'll see what comes out of that. But there's still a lot of open questions and therefore some uncertainty, and therefore a lot of what we're seeing is still human-in-the-loop type solutions, as opposed to just letting the generative AI loose on the end customer.

David Linthicum:

I was wondering about compliance. It sounds like we have more regulations that are emerging around utilization of this technology. But won't the technology itself allow us to manage compliance better with less of an effort, at least less of a human effort?

Ruben Falk:
There's a lot of activity in the area of being able to write regulatory reports, and all the documentation required by regulators potentially can be generated by generative AI. So, there's a lot of activity in that area. So, yeah, potentially it might be made easier by generative AI, but there's also going to be new regulations that have to be complied with. Some of the use cases only work if you don't have a human-in-the-loop. So, if you think of the use case of generating highly customized portfolio commentary for each of your individual investors, as a wealth manager advisor maybe of 100,000 individual investors that each have portfolios that are specific to them, and they could get a custom market commentary piece of output every day potentially, and that could be generated by generative AI. But that only works if there's no human in the loop. Otherwise, it becomes sort intractable and not scalable. So, in those situations, we'll have to find a solution to compliance that everybody is comfortable time. That might take a little time, but it's clearly what people are thinking is coming.

David Linthicum:
So, what about your customer's customer, so users of financial services, perhaps banking systems, things like that? What changes are they going to see that's going to be trickling down from utilization of this technology?

Ruben Falk:
I think as an end user you should be able to get better products, better customer services, and to some extent it should be transparent in the sense that as a customer calling into a call center or a customer receiving a fully customized market commentary, you're not really going to notice that it's generative AI that's in play, but you're just going to have a better customer experience. For that reason, it's going to be a competitive tool of the various providers. There may be some cases where, again, the solution filters all the way down to the end user.

For instance, again, in wealth management we have fintechs that are working on solutions around natural language search. So, you can say something like, "I would like to invest in a sustainable energy company or sustainable energy companies that don't use raw materials mined in a certain region." So, those types of solutions might become subject to regulatory and compliance comfort, might actually be exposed so that end users will have more of that generative AI experience in their own right.

David Linthicum:
So, final question, but I think it's important. What about ethical considerations? How does that factor into all this?

Ruben Falk:
It clearly factors in. Our philosophy is that we want to make existing job functions more enriched in the sense that repetitive tasks can be delegated to AI and generative AI, which leaves the employee to be able to pursue more enriching activities, where they can add more value with the help of human creativity and so forth. So, it's clearly a consideration, but we think look at the bright side. There's clearly a lot that can come out of sort of freeing up current human employees to do more enriching, value-add work.

David Linthicum:
I'm finding the same thing. There're way more benefits than any kind of deficits in terms of leveraging this technology if you look at the bigger picture. So, where do you learn about all this stuff? Where do you personally go to learn about financial services, generative AI, utilization of cloud, utilization of AWS?

Ruben Falk:
I read a lot. We have a lot of internal resources that I use. Also, I like Twitter [now known as X] for the purposes of following a lot of knowledgeable people in this space. So, with my Twitter [X] feed I get all of this new, exciting generative AI announcements pretty much in real-time.

David Linthicum:
I find social media is a great resource too because you follow smart people and they post what they're reading, and you kind of follow along and you get smart right along with them. So, where can we find more information about your service, you personally out on the web?

Ruben Falk:
Obviously, at AWS we have a lot of resources available that are easily Googleable. Myself, I'm on LinkedIn. My handle is my full name in one word. I publish a fair amount on my LinkedIn. We had a generative AI use case paper for financial services recently. I'm on various webinars and so forth, all of which can be found on my LinkedIn.

David Linthicum:
It's important stuff. You've got to remember that generative AI is kind of a revolution within a revolution. So, in other words, you start in the cloud and the ability to commoditize different services. Now we have the ability to leverage this core technology that is a true force multiplier for most of these businesses, and your ability to understand and learn them. How they apply in different industries is really important right now. The financial industry is probably going to benefit the most from this, in my opinion. 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, e-mail me at dlinthicum@deloitte.com. So, until next time, best of luck with your cloud journey. You guys stay safe. Cheers.

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