Patrick Moorhead: Hi, this is Pat Moorhead. Welcome back to day one of The Six Five Summit 2023. We are talking about the topic that everybody wants to know about, AI. Pretty much everybody wants to know about AI and I have the honor of having here someone from AWS, Archana who is going to tell us all things about AI at AWS. Welcome to the show.

Archana Vemulapalli: Thank you for having me, Pat.

Patrick Moorhead: Yeah, it's incredible. As an analyst, I always like to say that kind of pandemonium and change creates a lot of business for us, and I literally cannot go an advisory without talking AI now. I mean AI's relatively new, but not really, but we all got excited about this generative of AI. So anyways, welcome to the show. You have been making some really exciting updates over the past few weeks, which I have chronicled in my reports. Can you talk a little bit about those?

Archana Vemulapalli: Sure, happy to. So we've announced three different capabilities. The first we actually announced Amazon Bedrock, and this is a super easy way to scale and build generative AI applications using foundation models. And this service uses the best of breed foundation models, both from Amazon as well as partners such as Anthropic, Stability AI, as well as AI21 labs.

Then we also further built up on our instances, so we have Trainium and Inferencia, instances that give you the most cost-effective cloud infrastructure for generative AI.

And then last but not the least, we launched Code Whisperer, which is free for individual developers. Think of this as an AI coding companion that uses generative AI to really improve developer productivity. And one context to the comment you made earlier, these are early days for generative AI, but the crux of it is really building on the history of AI and ML, and AWS has a 20 plus year history, and that experience of delivering it at scale has enabled us to make sure we take care of the enterprise requirements as well as bring the best of breed capabilities to the customers.

Patrick Moorhead: Yeah, it is incredible the benefit that I think it gives AWS when multiple areas of your business is actually using artificial intelligence. And it is funny, I mean first if we want to get historical here, the first AI algorithms were developed in the 1960s.

Archana Vemulapalli: That's right.

Patrick Moorhead: ... And then we kind of went through this analytics AI stage, and then we went to machine learning and deep learning. And not that we got rid of any of the ones before, we're still actively using this, but here we are with generative AI. Why do you think generative AI has captured so much attentionally? Like what's different this time around? In fact, generative AI models were in use a few years ago. What changed?

Archana Vemulapalli: Yeah, I think you're absolutely right. There are a couple of things that have come together recently that have made this more immediate. One, is the ready availability of scalable compute capacity. This is super important.

The second, actually, is the massive proliferation of data that we see. And then both of these coupled with the advancement we're seeing in ML technologies have really led to all of the wave of innovation that we are seeing. I think, I mean generative AI has really caught the world by

storm. We're seeing consumer facing applications come up. We're showing the power of these ML models. And I'm sure Pat, you've worked with some of these models yourself and you see how exciting this space is. I think it's absolutely fascinating. Every customer we engage with, the art of the possible has just opened up and people are really looking to re-innovate, look at new models of growth, and I think we'll see a lot of good change come.

Patrick Moorhead: Yeah, it's exactly what it is. It's the art of the possible. I love that term. I mean, when you combine... When we double down on creating technology, we don't always know exactly the way it's being used, and it's kind of been a hallmark for AWS that really started off as for developers. And you put the right tools in the hand and the right power, at the right price point, with the right responsiveness, it's amazing what people will come up with. And I think we've all been blown away by generative AI with text, images, video, code, and a lot of other types of data. But I think five years from now we're going to stand back and say, "Wow, I don't think I expected that, but this is great." So part of your strategy is you have some open foundation models that are available and you also have some very finely tuned AWS foundation models.

Archana Vemulapalli: That's correct.

Patrick Moorhead: Can you talk about AWS foundation models and what they can be applied to?

Archana Vemulapalli: Yeah, I mean I think this is a super important area and you bring up a really good point. There are several foundation models available for implementing generative AI applications. Each has its own unique strengths and characteristics. And what customers really need to be thinking about is how do you leverage the choice of models to make sure you're solving for the right use cases? This is why, to your exact point, we believe in having a choice of cutting edge models. So like you said, we have the Amazon Titan models as well as the leading models from AI21 Labs, Anthropic and Stability. With our Titan models as an example, first we have a generative large language model, and this is for tasks such as summarization, text generation, classification, open-ended Q&A, information extraction, all the popular use cases you're seeing come up now is what this will address.

The second is an embeddings large language model. And so this translates your text inputs into numerical representations. And where this is powerful is for personalization and search. Similarly, when you talk about some of the third party models, we have, AI21 Labs for example, follows natural language instructions to generate text in multiple languages. So you take Spanish, French, German, Portuguese, Italian, Dutch, they cover it. Claude again Anthropic's large language model is doing a variety of conversational and text processing tasks. And this is based on a lot of work they've done in building responsible systems. So with Bedrock, we essentially bring all of these together.

The other one is Stability AI, we think about stable diffusion, the most popular of its kind, the capability of generating unique, realistic, high quality images, which you're seeing around a lot these days is again, another model there. So Bedrock again is the core essence of bringing these best of breed together. And really the goal for everyone is to think about what the key use cases are, and how you leverage these models to get to the outcomes you need.

Patrick Moorhead: Yeah. Again, I like the approach. I mean, there are some things that AWS and Amazon know well, I mean basically they're the template for what people should be doing and they want get access to it. And then there's open models to choose. I mean, who doesn't love choice? And really, if I've learned anything about researching at AWS for 12 years, a lot of it is about choice.

Archana Vemulapalli: Yeah, you know as well.

Patrick Moorhead: Yeah, you have your own and you bring the best from the outside. So some of these, while there's been a lot of talk, and videos, and viewpoints on the consumer point of view, and obviously you serve that through B2B to C. There's also B2B and models that can just raise the level of productivity, everybody in the organizations. And we've even seen some press releases from some companies saying, "Hey, I'm going to get this much productivity out of the team now." And the question is, let's talk about developers. What are you doing around developers to enhance that productivity? And like I said, AWS is for developers, it's how you started.

Archana Vemulapalli: So spot on, and this is one thing that has been front and center in all the conversations we've had from customers. Optimization is one of the key outcomes people are looking for from generative AI, and developer productivity is front and center in that conversation. So today you see software developers spending a significant amount of time writing code, and that's a very standard way in which it's approached. They spend a lot of time in what's changing in complexity, in what's changing in technology terms, and they're trying to drive a lot of these capabilities.

With Code Whisperer, the goal really is to give a coding companion that uses a foundational model under the hood. And then the goal again here is to rapidly improve productivity. And just to give you context, during the preview, we ran a productivity challenge and participants who used Code Whisperer completed their tasks 57% faster on average, and that's 27% more likely to complete them successfully than those that didn't use Code Whisperer.

And this is just one productivity challenge we ran. We also then had companies like Accenture use Code Whisperer for accelerating coding as part of their software engineering best practices. Using Code Whisperer, the company reduced development efforts by 30%. This is significant. As you think about it, you also want to help developers produce code responsibly, and Code Whisperer then also filters out code suggestions. So if there's anything that's considered biased, unfair, if there's security issues, there are a lot of things built in that help you produce better quality code in your organization. So for us, we believe this is a no-brainer with the value it brings and we really hope that this will be adopted in a large scale.

Patrick Moorhead: One of the things that first struck me about researching and writing about Code Whisper were the amount of languages and IDEs that were supported out of the box. And this is not fiction, this is GA, right?

Archana Vemulapalli: That is correct. It is GA.

Patrick Moorhead: That's the cool part about it and with more languages and IDEs coming out there. So it looks like a colossal amount of work. Sidebar here, my son is a data science major and I talked to him

about Code Whisper and somehow he didn't know was available, but he is like, "I am all over this."

Archana Vemulapalli: That's awesome.

Patrick Moorhead: I'm thinking, yeah, he was really excited about it. So he's going to do a test drive and let me know what he thinks about it, but if anything, what this does-

Archana Vemulapalli: It's going to change-

Patrick Moorhead: ... Yeah.

Archana Vemulapalli: I think it's going to change the developer experience plan in a fundamental way. And you'll see development and software development before and the way it used to be done and there's going to be a complete shift in the way it is going to be done in the future.

Patrick Moorhead: Totally. And by the way, the same arguments of calculators being bad for the classroom or stuff like that, it's like, no, no, no. We've seen this before. Better technology and for lack of a better term, it really democratizes-

Archana Vemulapalli: That's right.

Patrick Moorhead: ... A lot of these things for developers. So developers aside, I mean by the way, super exciting, to me it looks like magic, okay. But what does this mean from across the enterprise? What does this mean for information workers? What does this mean for maybe frontline workers? Is generative AI ready for enterprises? And on a broader basis?

Archana Vemulapalli: I do believe it's ready for the enterprise. And when you start to think about the capabilities one can bring to bear and the use cases that you start to pan out for, this is where there's massive opportunity. Customers have told us there are a few big things they need from generative AI within their businesses. They want one, an easy way to find and access high performing foundation models that give them outstanding results that meet their needs.

Second, if you go beyond the models, what customers want really is integration into their applications so that they can go and leverage it in a way that they're not managing a huge cluster of infrastructure. You know this one, right?

Then this one, and finally, they really also think about saying, "Hey, what we want to do, we want it to be easy to take a base FM and build differentiated apps using their own data."

So the reason I say the enterprise is ready is they want to differentiate. They're also equally worried about getting disrupted if they don't move fast. And so that appetite of wanting to disrupt, wanting to protect their business and IP, and accelerate where they can has really opened up the aperture of the use cases we're seeing from customers. And that coupled with the, I want something that's easy to use, I don't want to be in the business of managing infrastructure, and I want to build differentiated applications with my data quickly in a secure and protected, in a private way, Is what is going to drive adoption in a big way.

Patrick Moorhead: Yeah. I'm sure you've learned nothing else about being the pioneer in the public cloud. That simplicity was the name of the game upfront. And then over time, as people got experience with it, Hey, what knobs, what bells and whistles? What can I change as I move forward? And once you get them on Simplicity, I get, but how can I customize this? And so can you tell us a little bit more how do customers customize, and optimize inside of Bedrock, maybe with their own proprietary set of data or something different?

Archana Vemulapalli: So you are so good with that question because Simplicity is going to be one of the primary drivers in how people leverage this in a way that scales.

Patrick Moorhead: Yeah.

Archana Vemulapalli: One of the things we were very intentional about with Bedrock, with the service experience was that customers can get started quickly without any specific ML expertise. They can privately customize the foundation model with their own data and easily integrate and deploy them into the applications using data of use tools and capabilities that they're already familiar with. So this includes, again, SageMaker integrations like experiments. So you have to do all of, you can do all of this without having to manage the infrastructure, which is a big concern they have.

The other piece is, again, to customize a model also is easy. It's as easy as pointing Bedrock to a few labeled examples in S3. And then the service can fine tune the model for a particular task without you having to annotate historically large volumes of data, which they used to do. Examples, in fact, as few as a 20 should be enough. So we believe we've really put the time and effort into making this experience as seamless and easy for customers without them having to go, we build yet another workforce on this, but to leverage that, the strong developer capabilities they have in a good way.

Patrick Moorhead: No, I love that. I loved how you turned around customization. It's simple customization.

Archana Vemulapalli: That's right.

Patrick Moorhead: I got you. So one of the reasons that the public cloud and AWS has been so successful is, it's simple in a lot of different ways. And one of it, that they can swipe a credit card and get running almost immediately. Now, there's also CIOs that are like, "Oh my gosh, the ability to go out and do this, isn't this going to be a costly thing?" And then when you look at some of the research that has come out, it could cost 10X to train a generative AI model than it does let's say a standard run-of-the-mill machine learning.

Archana Vemulapalli: That's right.

Patrick Moorhead: ... A deep learning model. And I'm curious, how do customers manage costs when they're building, running, or inferring and customizing these foundation models?

Archana Vemulapalli: I think it comes back to the point you raised. You can drive innovation when you're doing innovation at scale, when you can do it simply, and cost effectively. And that's been the core belief system in Amazon and in AWS. And so here, very simply, when whatever customers are

trying to do with their foundation models, fundamentally what they need is the most performant cost-effective infrastructure. This is why over the last five years, we have been very methodical about investing in our own silicon to push the envelope on performance, and price performance for demanding workloads you know that ML training, and inference causes.

With AWS via Trainium, as well as AWS Inferentia, these chips have the lowest cost for training models and running inference in the cloud. So as context for Trainium, we can deliver up to 50% savings on training costs over any other EC2 instances.

So as you start to think about the future, when foundation models are deployed at scale, in an enterprise, across an organization, more costs are going to be associated with running models and doing inference. Similarly for inference are inference AI instances delivered up to 4X higher throughput and that's a 10X lower latency compared to prior generation inference AI instances. Again, these capabilities are at 40% better inference price performance than other comparable Amazon EC2 instances, and the lowest cost for inference in the cloud and you know inference is where the cost is. And we have customers that are already seeing benefits. We have a customer like Runway that's seeing 2X higher throughput within NP2 than comparable Amazon EC2 instances for some of their models.

So we are extremely optimistic that we have really focused on the cost parameter and we've looked at it from a scale lens, and we really hope that this will come to bear in a good way for a lot of our customers.

Patrick Moorhead: Yeah, I've been research researching your custom silicon for years now. I spent 10 years at a merchant silicon provider, so I'm looking at it a slightly different lens, but I'm very impressed with the capabilities, and when it comes to even Inferentia you're using it internally on one of the largest models that that's out there internally. So I really like that you use your own silicon internally and you're very rapidly getting more and more customers every day up on that. And-

Archana Vemulapalli: That's right.

Patrick Moorhead: ... Hey, if you want merchant silicon from, 'insert your name here,' you offer that too. And I had a great conversation with Dave Brown, which is also airing today, on day one, kind of looking historically and talking about the latest and greatest, and what people should look at in the future. So I think it is... And again, it's not just a story, it's what you're building, but it does make a pretty good story.

And I like to call this full stack AI all the way down to the silicon, all the way up to the developer tools.

Archana Vemulapalli: That is correct.

Patrick Moorhead: And in some cases the applications that go on and everything in between. It's really kind of where you want to start. If you're more comfortable doing models yourself, you've got a layer for that. You've got easy buttons as well. And I really think that's what enterprise IT is looking for because everybody has a different story. A lot of the born in the cloud folks, they know how to do this. It might be born in the cloud folks who might say, "Hey, I don't want to actually be really, really good at this, or there's no way I can bring in that cluster." Like you had talked about

before. And managing a cluster, keeping it up, particularly when it's full of hardware, it is hard. It's a challenge, that I'm hearing because the exponential size of this is daunting.

Archana Vemulapalli: I think that's spot on Pat, I fully agree with you, and that's what we're excited about. I think we genuinely want to make this an environment where people have the ability to innovate at scale, and where people don't hesitate about innovating at scale because of cost concerns. And so all of our investments and our strategy has been aligned with that direction. So we're just super excited to see what organizations are going to do partnering with us. I think it's going to be an amazing space in the coming years.

Patrick Moorhead: Yeah, I'm super excited about it. Thank you so much for coming on The Six Five and being part of The Six Five Summit, and quite frankly, the track that has the most interest. We signed up speakers in very quickly. I mean, I guess I shouldn't be surprised because it is, it's part of every conversation now. And when I go to events, when I talk to enterprises, the first thing they want to hit me up on is, "Hey Pat, tell me about this generative AI thing. What should I do with it?" So thank you so much.

Archana Vemulapalli: My pleasure. Thank you again.

Patrick Moorhead: So this is Pat Moorhead signing out here for a day one track here in AI. Stay tuned. There's more AI action around and hang out for more day one, where we're talking cloud infrastructure as well, among other things. And we have day two and three. The great thing is, if you can't sit in front of your PC, or your front of your tablet for three straight days, it is on demand. You pick what sector, what speaker you want, we can go from there. So for that, thank you. Good morning, good evening, good afternoon, wherever you are on the planet. Thank you.

Archana Vemulapalli: Thank you.