Amazon Web Services CEO's Keynote At AWS re:Invent 2019

Company Participants

- Andrew R. Jassy, CEO
- Matt Wood, VP of Artificial Intelligence

Other Participants

- David Brent Shafer, Chairman & CEO, Cerner Corporation
- David Michael Solomon, Chairman & CEO, The Goldman Sachs Group, Inc.
- Hans E. Vestberg, Chairman & CEO, Verizon Communications Inc.

Presentation

Andrew R. Jassy {BIO 15111610 <GO>}

Thank you. Hello and welcome to the 8th Annual AWS re:Invent. It is awesome to be here. This is our favorite time of the year. You're here with 65,000 of your peers. And as you know, AWS re:Invent is not a sales and marketing conference, it's a learning and education conference. And so everybody's favorite part of the week are the breakout sessions. This year, it will be no different. We have over 3,000 of them, with most of them being led by partners or customers. So you can get the real scoop about the platform. But I am going to need every minute of these 3 hours we have scheduled. I have a lot for you, including some things that are good at the end. So I'm going to get right to it and giddy up.

So when we were thinking about throwing a conference back in 2012, we spent a lot of time debating what the name of the conference should be. And you can decide if you like what we chose or not. But it was very intentional. And the reason we chose this name was it reflected what we were seeing at the time, which is it was incredible the number of large and small companies alike and the way that they were inventing and the pace that they were inventing and how many large enterprises were completely reinventing their customer experience, all with AWS in the cloud as the lynchpin. And you can see it if you look at the millions of active customers that are using the platform today.

And if you look at what's happened over the last six years or so, the enterprise has very dramatically changed their view of the cloud and have adopted AWS in the cloud in every imaginable vertical business segment: so in financial services, you see it with Capital One, in Goldman Sachs, in Barclays, in HSBC, in Intuit, in the Commonwealth Bank of Australia; in life sciences and health care, you see it with Merck, with Pfizer, in Bristol-Myers Squibb, Johnson & Johnson, Novartis, AstraZeneca; in manufacturing, with GE and Schneider Electric and Siemens and Philips; in energy, with Shell and BP and Hess and Halliburton. Every imaginable vertical business segment is using AWS in the cloud in a very meaningful way. And you also see it in the public sector where we have 7,000 government agencies worldwide using AWS, 10,000 academic institutions and over 25,000 nonprofits. So very broad adoption. And the reason that these bigger companies have started moving to the cloud is because what's happened over the last 12 to 13 years is that start-ups have completely disrupted long-standing industries, largely from a standing start on top of AWS.

This is often a time where I reel off a number of start-ups that are using the platform. And most of the successful start-ups over the last 13 years have built their businesses on AWS. But I'll give you a few examples that I think just illustrate what I'm talking about in terms of this revolution.

Think about in the old days when you needed a taxi. And you called a taxi company and the dispatcher would sometimes pick it up. And sometimes let it ring forever. And sometimes, they'd be nice to you on the phone. And sometimes, they'd be rude to you on the phone. And sometimes, the taxi will come when they said it was. You never knew when it was going to come. You get in the car, it will be a little bit filthy. Think about that experience today and how completely revolutionized it is by companies like Lyft and Uber and Grab and Ola and Careem, totally different.

Or look at accommodations. When you wanted to stay somewhere out of town, you used to always just default to a hotel. But look at how Airbnb has completely changed that industry. They have 4 million active listings at any one time and 2 million people a night staying in their rentals. It's just crazy.

Think about exercise. For those that used to use the bicycle for exercise. Yes. You can go out and ride a bicycle if you felt like dealing with the elements. You could then go to a spin class. But then you have to get in the car and go there and you have to go to a particular time. You could have a bicycle at home. But that's a little bit boring to exercise yourself. And think about, again, how Peloton has used technology to completely change that experience. Every day, they have 20 livestreams. They have 10,000 on-demand streams that you can look at. It's just totally changed the way that you exercise.

Then think about delivery of food. In the old days, unless you lived in New York City, the only thing you got delivered to you was pizza. But now with companies like DoorDash, Grubhub and Postmates, you can get virtually anything delivered to you. Believe me, I know this, my son is one of their best customers.

And so if you think about it, these start-ups have disrupted long-standing industries that had been around for a long time and really from a standing start. And it's why when we thought about what we should choose as a theme for the keynote today, we kept coming back to the same topic, which is the question and the topic we talk most about with companies, which is how should we think about transforming ourselves. How can we reinvent our business and our customer experience so we can be meaningful and sustainable over a long period of time. And transformation can mean transforming yourself, or it can mean transforming to meet new technology situations or opportunities.

And so this transformation question was the #1 question. And so we thought we would share with you what we think are 6 of the most critical components if you're going to make a significant transformation. But there's a big change and a big transformation that you have to make. You don't want to procrastinate. It doesn't get easier if you wait. In fact, the ditch gets deeper. And so when you think about the type of the first element of a big type of transformation like this, it turns out it's not technical. It's all about leadership.

And so when we look at the companies that make this transformation successfully versus those that just talk about it, there are 4 differentiators. The first is you actually need to figure out how to get your senior team aligned that they're going to make this change. It's not easy to make a big shift like this. And inertia is a very powerful thing. And it's easy to block in various parts of the organization, sometimes for well-intended reasons and sometimes for self-interested reasons. But it's easy to block. And if you don't find a way to have that senior management conviction and alignment that you're going to make that change and find a mechanism to get the issues on the table so you know you're making progress and you don't go several months down the road thinking you're making progress when you're not, you won't make that change. So you need that senior level alignment.

The second thing you need right alongside that is you need an aggressive top-down goal that forces the organization to move faster than it organically otherwise would. And let me give you 2 counter examples of this point. So if you think about several years ago, the CIO of GE, Jamie Miller, decided that she thought it was critical that GE move to the cloud to move much more quickly. And she got her top technical leaders together and said, "We're going to move 50 applications to AWS in the next 30 days." And she said for 45 minutes they told her what a dumb idea that was and how it would never work. And she listened to them very patiently and said, "I hear you. But we're going to do it. So let's go." And they got to about 42 applications in 30 days. But along the way, they figured out their security model, their governance model, their compliance model and they had success and built momentum. And all of a sudden, all the ideas started flowing in on what else they can move. And they're now about 3/4 of the way through moving several thousand applications to AWS, a second aggressive top-down goal that wouldn't have happened if they didn't set that goal that forced the company to move.

Now let me give you a counter example. I went to go see a company in the life sciences space. And it turned out that I knew the CIO through a friend of a friend. I've never met him before. But we had this connection. And I got to the meeting. And they were a little bit late coming out to see us. And they were in with their CEO. And the CIO came out and said, "I'm going to grab you." I think because he felt bad because we had this friend of a friend connection. And he said, "Before my infrastructure leader gets in the room, tell me how we're doing together." And I said, "Well we're doing fine. But you're not doing very much with us." He said, "Oh, it's definitely not true. Like I know we're doing a lot. I heard we've got all kinds of workloads running. We're experimenting. We're kicking the tires." I said, "Well you're using 3 EC2 instances." He said, "Oh, that can't be right." Then his infrastructure leader came in the room. And he said, "(John)," he said, "Are we doing a lot with AWS?" He said, "Oh, yes, we're doing tons. We're kicking the tires. We're experimenting with it, lots of things going." He said, "How many EC2 instances are you using?" And he said, "Like 3 or 4." And he said, "That's not what I mean by doing a lot." But it's easy to go a long period of time dipping your toe in the water if you don't have an aggressive goal that lets the organization know that it's a priority to make this transformation. By the way, that CIO went away 2 weeks later, set an aggressive top-down goal. And they're one of our top 5 life sciences customers today. But it needs that top-down aggressive goal.

The third thing is you got to train people. Lots of times, you have these conversations around a table, senior people get excited, they decide to move to the cloud. They come back to their companies. And they say, "Good news. Here's the cloud." And nobody has any experience using it. Now it's not that hard to use the cloud. But it takes a little bit of training. So that's why we train hundreds of thousands of customers every year.

Then the fourth thing that's important to do is that you got to make sure that you don't allow the organization to get paralyzed if you haven't been able to move -- figure out how to move every last workload. What we do a lot with our customers is we will go and do a portfolio analysis where we will go through all of their applications with them. And we'll classify them into the applications that are easy to move, medium hard to move, should go last because they have the most dependency and legacy, those that can easily be lifted and shifted and those that should be rearchitected before they move to the cloud. And we build them a thoughtful, methodical, multiyear plan to migrate. And what companies find almost always is that so many workloads are relatively easily able to move to the cloud and get all those benefits. And in fact, they inform a lot of the later workloads that are the hardest ones to move. So you've got to make sure that you don't get paralyzed by that.

So this first step of the transformation is not technical. It's very much about leadership. And it's about making sure you have senior level alignment, an aggressive top-down goal that forces the organization to move faster than it otherwise would, the right training and then a thoughtful, methodical, multiyear plan to make that migration. Now what you find is once, as a company, you make that decision that you're going to make this transformation and move to the cloud, your developers are raring to go. You are ready to go. And you want the broadest possible capabilities. And you don't want to be slowed down.

Last year, we talked about "I want it all and I want it now." This year, we're talking about "Don't stop me now. I'm having such a good time. I don't want to stop at all." And once you decide as a company that you're going to make this transformation to the cloud, your developers want to be able to move as fast as possible. They don't want to be constrained. They want all the capabilities they need to move everything that they want to move and to build anything they can imagine. And so they want the broadest possible capabilities. And there's nobody who has the capabilities that AWS has. We have over 175 services. Nobody is close to the same number of services. But it's not just the number of services, it's the depth of features and capabilities within these services.

And there's a lot of noise out there. And there are a lot of companies who've become pretty good at being checkbox heroes where they kind of look at something we have. And they rush to have it out there and say, "We have it, too." But when you look at the depth and the details of the offerings, they're pretty different. And you'll see this across all of these major infrastructure components: in compute, in storage, in database, in analytics, in machine learning, in IoT, in robotics, in messaging, in content distribution, in marketplace, in people services, very big differences.

And this is a quote that you see, we hear this a lot. This one happens to be from Expedia. But what we typically hear from customers when we talk to them is that they think we're a couple of years ahead both in functionality and with regard to maturity. And so instances, containers, network. I'm not going to spend a lot of time on the networking piece, in part because I'm going to let you be surprised by the press release we have coming out later today. And also, Dave Brown is going to reveal a lot of our -- the details of our new networking features in his networking presentation on Wednesday.

But I'll say a couple of things. First of all, you don't have compute without a great network attached to it. And you won't find a network that has more functionality and more capability than AWS is. If you look at our footprint of POPs and abilities to get on our backbone, we have a much broader footprint than you'll find elsewhere. It also is a network that has more places for you to do Direct Connect between your on-premises data centers and AWS. It's the only one to have 100 gigabit per second for standard instances. Then you also have the most capable network hub, what we call our transit gateway, which allows you to connect your on-premises data centers with AWS and then also set up across multiple AWS VPCs in different regions. It has the ability to take more connections. It has much more throughput than you'll find anywhere else. It has the ability to connect your branch offices more easily to AWS than you'll find elsewhere, the branch offices between each other with SD-WAN integration. And even multicast IP, which nobody else has, which we're just launching today. So a much more broadly capable network than you'll find elsewhere.

But I'm going to focus most of my comments on compute, on instances and containers. And so if you look at instances, to start, it's not just that we have meaningfully more instances than anybody else. But it's also that we've got a lot more powerful capabilities within each of those instances. We have the most powerful GPU machine learning trading instances, the most powerful GPU graphics rendering instances, the largest in-memory instances for SAP workloads with 24 terabytes, the fastest processors in the cloud with the z1d. You've got the only standard instances at 100 gigabits per second network connectivity, the only instances that have all the processor choices from Intel and AMD and ARM base, very different set of capabilities on the instances side. And if you look at the pace of innovation in AWS, on the number of instances that we've built, it's totally accelerated in a very significant way. We have 4x more instance types today than we did two years ago. And there are a couple of reasons why we've been able to innovate in a much faster rate.

The first is that we have spent a significant amount of time over a couple of years totally rehauling and reinventing our virtualization hypervisor layer. And if you -- we built it with a system that we call Nitro. And if you look at what Nitro does, it takes the virtualization of the security and the networking and the storage off of the main server where the lightweight hypervisor and the customer instances are and gives you back all that CPU was consuming before, which means that you get performance indistinguishable from bare metal at a much lower cost. It also means, because we've taken all these pieces off that main server and put them on Nitro chips that we built, that we can innovate in a much quicker fashion because we don't have to -- every time we make a change to one of those pieces, we don't have to have all of them changed in lockstep. And so it's why we are able to add our network optimized instances, for instance. So quickly, which you won't find elsewhere because we've separated those pieces into separable bunches.

The third thing is that we have a security capability in Nitro that I also think is meaningfully better for you, which is most of the traditional hypervisors have a trusted domain, which people often call Dom0. And they're for things like being able to add VMs and agents and troubleshooting. And you have a limited number of people as a provider that you allow to have access to that trusted domain because it's got all the customer instances on it. You have to be careful to lock it down, et cetera. With Nitro, because we've moved the security off that main server into a separable Nitro chip, it means that we just lock down that main server with the customer instances. No one can access it. We don't have to worry about the control of that, it's just not accessible, which is much better security posture for you.

So the first thing was we totally overhauled and reinvented the virtualization layer with something called Nitro, which has helped us innovate at a much faster rate. And you'll see that throughout this conversation.

Second thing is that we decided that we were going to design and build chips. And I think in the history of AWS, a big turning point for us was when we acquired Annapurna labs, which was a group of very talented and expert chip designers and builders in Israel. And we decided that we were going to actually design and build chips to try to give you more capabilities. And I think that while lots of companies, including ourselves, have been working with x86 processors for a long time. And Intel is a very close partner and we've increasingly started using AMD as well, if we wanted to push the price performance envelope for you, it meant that we had to do some innovating ourselves. And so we took this Annapurna team. And we set them loose on a couple chips that we wanted to build that we thought could provide meaningful differentiation in terms of performance for you on things that really mattered and that we thought people are really doing it in a broad way.

And so the first chip that they started working on was an ARM-based chip that we called our Graviton chip, which we announced last year as part of our A1 instances, which were the first ARM-based instances in the cloud. And these were designed to be used for scale-out workloads. So containerized microservices and web tier apps and things like that. And we had 3 questions we were wondering about when we launched the A1 instances. The first was will anybody use them. The second was will the partner ecosystem step up and support the tool chain required for people to use ARM-based instances. And the third was can we innovate enough on this first version of this Graviton chip to allow you to use ARM-based chips for much broader array of workloads.

And so on the first 2 questions, we've been really pleasantly surprised. You can see this on the slide, the number of logos, loads of customers are using the A1 instances in a way that we hadn't anticipated. And the partner ecosystem has really stepped up and supported ARM-based instances in a very significant way. The third question on whether we could really innovate enough on this chip, we just weren't sure about. And it's part of the reason why we started working a couple of years ago on the second version of Graviton even while we were building the first version because we just don't know if we're going to be able to do it and it might take a while.

And so I'm excited to announce today the launch of a new set of instances: the M6g, the R6g and the C6g instances for EC2, which is a new generation of ARM-based instances powered by AWS Graviton 2. So these are pretty exciting. And they provide a pretty significant difference over the first version of the Graviton chips. Each of these have 64-bit customized cores with AWS-designed 7-nanometer silicon. All the instances have up to 64 vCPUs, 25 gigabits per second of enhanced networking and 18 gigabits per second of EBS bandwidth. They have -- versus the first Graviton chip, they have 4x more compute cores, 5x faster memory and overall 7x better performance than the first Graviton chip. But arguably, most importantly, they have 40% better price performance than the latest generation of x86 processors. That's unbelievable if you think about that.

So we're very excited to give these to you today. The M6g are available today. The R6g and the C6g will be available in early 2020. So what we also did was, simultaneously, we split off a piece of the Annapurna team to build a second chip. And again, we are trying to pick things that could totally change the game for you with regard to what you can use. We started with these Graviton 2 chips that now, with that 40% better price performance than what you can get on the latest generation of x86, you can run virtually all your workloads on them. That's a huge game changer for you all.

But then we also started working on something that we thought would also be a game changer. And that's really around machine learning. And we've talked a lot as a group over a few years about training with machine learning. It gets a lot of the attention. And they're hefty loads. And we have the instances that perform the best and are most powerful on machine learning trading with our P3 and P3D instances. But if you do a lot of machine learning at scale and in production like we have. And a lot of you inside the room have done, you know that the majority of your costs is actually in the predictions or the inference.

And just think about an example, I'll take Alexa as an example. We train that model a couple of times a week. It's a big old model. But think about how many devices we have everywhere that are making inferences and predictions every minute, about 90% -- 80% to 90% of the cost is actually in the predictions. And so this is why we want to try and work on this problem. Everybody is talking about training. But nobody is actually working on optimizing the largest cost for you all with machine learning.

And so we announced last year in my keynote that we are working on an inference optimized chip called Inferentia. And I'm excited to announce today the launch of the Inf1 instances for EC2, which are backed by our new Inferentia chips. And so our Inf1 instances have a lot of things to be excited about. It will have low latency. It'll have 3x higher throughput. Just think about that again, 3x faster throughput, 40% less cost than the current best inference instances that are in video chips, just a significant innovation from this team. 2,000 tera operations per second. We've integrated with all the major frameworks, with TensorFlow and with MXNet and with PyTorch. And it's available today in EC2. But we'll make it available for SageMaker and ECS and EKS in early 2020.

So when you think about instances, in addition to having the most number of instances and then the most powerful instances within each of those categories and then the capability with Nitro to innovate at a much faster clip, when you layer on top of that our desire and willingness to design and build chips, it gives you capabilities in the instances space unlike you'll find anywhere else.

But that's instances. Let's talk about containers a second. So increasingly, customers are using containers for all kinds of workloads both because of the advent of microservices architectures and also because it makes it quicker and faster for people to deploy. But this is another area when you think about containers that we have a lot more capabilities and a lot more resonance than elsewhere. If you look at all the containers in the cloud, 81% of them run in AWS. And that's a significant part because we just have a lot more capabilities. So all the other providers effectively have 1 container service, which is a managed Kubernetes service. We have 3.

So when we started building container services back in 2014 before there was really a very popular orchestration engine, what customers said to us was, "we don't just want a container, we want a container as deeply integrated with the rest of the AWS platform capabilities." And so we built something called Elastic Container Service, or ECS, which has grown unbelievably fast and continues to weed a lot of customers like Verizon and GoPro and Fox and McDonald's, Duolingo that use these. But because we control the development of ECS -- with a lot of help and a lot of input from all of you, thank you very much and keep it coming, please. Because we control the development of ECS, it means that we can launch new features where they integrate with ECS right from the get-go. So if customers want the most deeply integrated container service out there, they choose ECS. But not surprisingly, as

Kubernetes became very popular, lots of customers wanted us to do something there. And if you look today, 84% of the Kubernetes that runs in the cloud runs on top of AWS. So we have a lot of Kubernetes customers. But they understandably wanted a managed service there. So we built the Elastic Kubernetes Service, or EKS, which has grown like a weed as well since we launched it a couple of years ago, really, really fast.

And so customers say, "Well it's awesome that I have these choices of 2 managed container services in ECS and EKS. But I would prefer not to have to worry about servers or clusters. I want to manage containers at the task level." And so that's why we built something that we launched a couple of years ago called AWS Fargate. And Fargate is a serverless container offering. It's the only offering like it anywhere out there. And all you do is you tell us the CPU and the memory that you want, you upload the container image and then Fargate does all the rest for you. It deploys it. It rightsizes the compute. No servers, no clusters, no provisioning for you. And if you look at the popularity of Fargate, it kind of blows us away. We thought people would be interested in it. But I don't think we anticipated it will be this broadly demanded. For new customers this year, new container customers in AWS, 40% of them start with Fargate. And that's because it's so much easier to run containers that way.

Now we launched Fargate with support for ECS because, again, since we control the development of ECS, we didn't have to coordinate with anybody. It was much easier to do so. And trying to make it work for Kubernetes has not been easy. And customers, understandably, who use Kubernetes said, "Well we love the idea of Fargate. But why won't you make it work for Kubernetes." So I'm excited to change that for you right now with the launch of Amazon Fargate for Amazon EKS. So now our Kubernetes customers are able to get all the same serverless benefits of running containers in AWS. And so what that means now is we have 4 container offerings for you to choose from. For those that like to manage at the server and cluster level and you want that flexibility to stitch things together, you can use either ECS or EKS. And for those that want to operate at the task level and not have to worry about servers and clusters, you can use Fargate for either ECS or EKS. So very exciting.

So when you make this decision that you're going to transform yourselves and you're going to move to the cloud. And you get that senior level alignment and you set that aggressive top-down goal that force the org to move, your developers want to go. They don't want to be held back. They want the platform with the most capabilities, not a fraction of the capabilities, the most capabilities, especially because you don't have to pay for it upfront.

And if you look across the platform, this is the bar for what people want. If you look at compute, they want the most number of instances, the most powerful machine learning training instances, the most powerful machine learning inference instances, the most powerful GPU rendering instances, the biggest in-memory instances for SAP workloads, 100 gigabit per second connectivity with standard instances, the access to all the different processor options. They want not just one container, they want multiple containers at both the managed level as well as the serverless level. Then they want the network with the most capabilities, the most functionality, the broadest footprint, the best performance and then the capabilities with things like transit gateway that make it much easier to set up your global network. That is the bar for what people want with compute. And the only ones that can give you that are AWS. And it makes it, by the way. So much easier, not only to move all your existing workloads over but also to allow your developers to build anything they can imagine with the right tools at the right price as quickly as possible.

Now you may have noticed as you are walking in that we had some DJs playing music, as we often do, before the keynote. And we had -- this year, we had 2 DJs. We had a woman DJ and a male DJ who did the last 15 minutes. And I'm going to actually bring up to the stage the male DJ, whose DJ name is DJ D-Sol. And he's actually not going to come up and talk to you about music, although he actually have very interesting things to say about music, I enjoyed his selection tape. It turns out that, that DJ is a CEO of Goldman Sachs. And -- it's true, yes. Yes, he's awfully good at DJ-ing in addition to being a CEO. And so I'm going to ask David Solomon to come to the stage and share with you the transformation that Goldman Sachs is making using AWS in the cloud. David?

David Michael Solomon {BIO 1503614 <GO>}

Thank you, Andy. Thank you for having me. Good morning, everybody. The one thing I'll tell you, for sure, the most fun I will have today was 15 minutes from 7:45 to 8:00 a.m. I'm really happy to be here. I'm standing up here looking at all of you, some of the world's best technologists at one of the biggest technology conferences of the year. And I know you're wondering, what's the head of a bank doing here? Well the world is changing. You just saw the CEO of Goldman Sachs DJ in Las Vegas. And not for the first time I would say. But let me tell you a little bit about how we're using cloud technology to change our business.

Goldman Sachs is a financial services company. We provide advice, we lend and invest money, we make markets and we manage risk. We stand in the middle of a market with trillions of dollars of flows. We advise on hundreds of M&A transactions a year and underwrite millions of stock and bond offerings. Ultimately, we help our clients achieve their financial goals. We care deeply about our clients' goals because they aren't intangible, abstract things, they are real-world problems. And everything we do from simple lending to complex derivative trading serves a real purpose like helping you renovate your kitchen or finance your expanding contracting business.

Finance is as simple as that, or it should be. The reality of it is finance is much more complex. And Goldman Sachs' job is to try to simplify it, to make it as easy, intuitive and as effective as possible. How do we do it? In-house, we have around 9,000 engineers who make up about 1/4 of our 38,000-person workforce. We have some of the best engineers in the world working on some of the most interesting problems. And everything we are building is trying to make finance work better. Cloud technology allows us to do our job in a way that's simple, all while accounting for the complexity of our industry and helping us ensure that our work is safe, secure and responsible.

While we work with a few cloud providers, AWS was the first because of their immense capabilities and the astounding pace of their innovation. A few years ago, my colleague, Roy Joseph, was on this stage here to talk about how we work together to develop the Bring Your Own Key solution for AWS' Key Management Services, a crucial data privacy development that's allowed us to fully, as an organization, embrace the cloud. Not to be too on the nose. But AWS' cutting-edge approach to technology really unlocked something for us here.

Since then, we've been busy, with the help of AWS, building predominantly cloud-based businesses. I want to talk to you about a couple of those for a few minutes here today. The first is our credit card business. In conversations with our clients, we realized that the credit card could be a much simpler thing. And it could provide consumers with a new way to relate to their spending, repositioning the credit card as a tool that was truly on your side. Building the credit card platform that underpins Apple Card took a number of our engineers as long as -- along with a very strong partnership with Apple, Mastercard and, of course, AWS. It would have made sense for us to get into the business if we had to maintain fields and fields of on-prem data centers to do it. The only reason we were able to deliver these capabilities digitally and at scale is because of cloud technology.

Apple Card launched just a few months ago. And it's already one of the most successful credit card launches ever. It's part of Marcus, our digital consumer banking business, which today is just three years old. Although we're 150-year-old company, we're still new to consumer finance. But I think it's pretty clear that we're on to something. Today, through Marcus, we have $55 billion in retail deposits and millions of clients. We're putting cloud technology to work in other areas of our business as well. Big companies make trillions of dollars of payments every day. But right now, this space is dominated by legacy architectures, manual processes and slow turnaround times.

Next year, we'll launch our transaction banking service, a digital platform that helps corporations manage their cash built entirely on a cloud-native stack provided by AWS. Goldman Sachs is already using it to make billions of dollars of our own payments in 5 currencies every day, providing us enhanced transparency, trackings around our cash flows while saving us a lot of time and money in the process. We're eager for our clients to experience the service and look forward to sharing more details about our rollout plan next year in 2020.

Historically, financial technology has been powerful and fast. But it's lagged behind consumer and high-tech in terms of elegance and simplicity. If you want a bank account, you have to wait for your funds to clear. If you want a loan, you have to wait for approvals. If you would want advice for your company how to better manage your balance sheet, you have to send an e-mail and ask for a meeting. It doesn't have to be this way. We can do better. But we're not there yet.

Finance is the perfect place to take new technologies and have an immediate real-world impact. Our data scientists have been using AI and machine learning techniques for years. And we're already pushing the research community to consider what's possible when you apply quantum power -- quantum computing power to financial problems. Well Goldman Sachs, there's corporations, governments, institutions and individuals, we're also building for developers. The same way you go to AWS for their best-in-class cloud services, we want to be your first choice to provide services that enable you to build financial functionality directly into your applications and workflows.

Already for our institutional clients, we're making the capabilities of our powerful securities database available directly through a platform called Goldman Sachs Marquee. The real power of Goldman Sachs Marquee lies in the scalable services you can access directly through our APIs. We've published some of our APIs on developer.gs.com. And we will continue to add more over time. Our clients already use the power of AWS to access a number of these services. We're migrating production of Marquee to AWS. And starting next year, we will be delivering new products and services to our clients there directly.

My goal is for Goldman Sachs to lead the way in building financial services technology. And we're going to succeed in no small part during the work we have done and will continue to do with Amazon Web Services. Still, our job will remain the same: to assume the burden of complexity and make finances easy, intuitive and as effective as possible. We want to enable you to focus on building things that we couldn't dream of, things that change the way the world works for the better because, sure, finance is complex. But it should never be a drag on innovation. Like AWS has been for Goldman Sachs, finance should only ever be a business accelerant.

I look forward to seeing what we can build together. Thank you, all very much for having me. Have a great day.

Andrew R. Jassy {BIO 15111610 <GO>}

Thanks, David. Really, it's an honor for us to be partnering with Goldman Sachs. We're really excited about what we're doing together. So thank you for all the partnership.

So when you're making this transformation and you make the decision you're going to make this big move and then you give your developers the access to all the tools to let them make this change quickly and flexibly, there's another thing you got to think through, which is you got to think about what are the things you're going to take with you and what are the things you're going to leave behind when you're making this big move. It's what people often call modernization. That's a question that companies have been asking about their on-premises infrastructure for many years.

When they compare it to the cloud and it costs more capital; it's more expensive on a variable basis; you don't get the elasticity; it doesn't have near the capabilities of the cloud; it's launching 2,500 new services and features a year; you have to spend your scarce resources, which are engineers, on the things that are undifferentiated heavy lifting; and you don't have the same security posture, when you objectively look at whether you want to toil away at refining your on-premises infrastructure for the next several years, most companies are saying, "No. Thanks. If that's moving up, then I'm moving out."

And so when you make that decision that you're going to modernize, you have to make a lot of decisions and you have to decide for yourself what you're going to bring. And it's a little bit like moving from a home. Look, there's a mainframe. It also looks like there's some audit notices and some licensing changes and pricing increases. Well when you're going to move, you have to decide what you're going to actually take with you and what you're going to leave behind. And it turns out when companies are making this big transformation, what we see is that all bets are off. They reconsider everything. And so I thought I would share some of the biggest decisions and some of the trends we see in modernization from companies that are making this big transformation.

So the first thing is companies are trying to move away from mainframes as quickly as they can. Every industry has lots of companies with mainframes. But people want to move away from it because they're expensive and they're slow and they're complicated. And hardly anybody has mainframe skills anymore. In fact, one of the stories I like is one of our enterprise customers was making a big shift in moving its mainframe and teasing it apart and moving it to AWS. And they got to the very last step where they needed to find the credentials to decompile something. And they couldn't find anybody in the company who have those credentials. And they realize that only 1 person had them. It was this woman who had retired 10 years ago, moved out of state, her name was (Ginger). And they were able to find her. And she gave them the credentials. And they made this last step.

But you can't always find Ginger. And so people don't want to be using mainframes over the long term. And we have lots of customers moving away from them. And we see a couple of different patterns. We see some companies that basically just tease the whole thing apart into microservices and then relaunch it in AWS. It's what Western Union did. We have some companies that are methodically picking certain workloads and moving them one by one away from the mainframe. So the mainframe becomes less central and kind of lives in a place with just nonessential tasks. And that's what companies like Vanguard and Allianz and the U.S. Air Force are doing. But make no mistake about it, companies are trying to move as quickly as they can away from mainframes and having a lot of success doing so.

Second thing we see people move and modernize from are those older guard commercial-grade relational databases. And we've talked about that here in this keynote for a few years. People are trying to move away from Oracle and SQL server because they're expensive, they're proprietary, they have high amounts of lock-in and the licensing terms are just downright punitive. And I think that we don't meet customers that aren't looking to flee Oracle. But one of the things that we've seen over the last couple of years as a change is people are trying to get away from SQL server pretty quickly as well. And one of the reasons is just you see this return to the ways of old from Microsoft where they're not prioritizing what matters to you guys and to customers.

So let me give you an example. For many years, you were able to take the SQL server licenses that you'd bought yourself and then bring your own license and run them where you want it to, in a relational database service as an example. One day, Microsoft decided that they didn't want to let you do that. Was it good for you? Hell, no. Is it good for Microsoft? Maybe. I think they think so. But people are sick and tired of being pawns in this game. And it's why they are moving as fast as they can to the open engines like MySQL and PostgreS and MariaDB. But to get the performance that you need from these open engines that compares well to these commercial-grade databases, it's hard. It takes a lot of work. We do a lot of it in Amazon. It's doable. But it's hard. And it's why all of you asked us to try and solve this issue, which was to give you a database option that works on the open engines but that has comparable performance to the commercial-grade databases.

And it's why we built Amazon Aurora, which we launched in 2015. And Aurora has versions that are completely compatible with PostgreS and with MySQL. It has several times faster performance than the typical high-end implementations in those community editions. It's at least as durable and performant and available as the commercial-grade databases. But 1/10 of the cost. And it's why it's been the fastest-growing service in the history of AWS since its inception in 2015. And you've got tens of thousands of customers who are using and moving to it. It's a very broad list like Bristol-Myers Squibb and Fannie Mae and Electronic Arts and AstraZeneca and Liberty Mutual and NASDAQ and Hulu and Verizon and FINRA, a very broad group. And it's pretty amazing to us how fast people are moving to Aurora at this pace.

Mainframes, relational databases. The third area we see a lot of modernization decisions is around moving from Windows to Linux. And this has been happening for several years anyway. IDC estimates that in 2020, about 80% of the workloads deployed will be Linux workloads. And Linux has grown about 4x faster than Windows. And there are a few reasons for this. First, people don't want to pay the tax anymore for Windows and particularly when they know the price goes up frequently. Second, it turns out, because there's such a vibrant community around Linux and a more vibrant community around it than anywhere else that all the features and all the security things happen much quicker there.

And third, again, people aren't so keen to have one owner of an operating system especially when they're prone to raise prices, which happens a lot, or just change the licensing terms. If you look at Windows in the cloud, 57% of Windows in the cloud runs on AWS. A company that owns an operating system maybe isn't so crazy about that. So they just decide to change the licensing rules for you. And they said, "Hi. new versions of Windows can run on dedicated instances in other cloud providers," trying to stranglehold those workloads back to their platform. Again, I don't think customers are so keen about having one company own that operating system they're building their business on, which is why you continue to see people moving from Windows to Linux.

The fourth area of modernization that we see is around the partners you choose. And so if you look at most ISVs and SaaS providers, they will adapt their technology infrastructure platform to work on a technology infrastructure platform. So they'll doubt their software will work on a technology infrastructure platform. Some will do 2, very few will do 3. And they all start with AWS just because we have such a significant market segment leadership position. And it's why you see a much more vibrant collection of ISVs and SaaS providers on AWS as you're moving to the cloud and want those capabilities. And you see it with -- Salesforce runs the vast majority of what they do on top of AWS as does Workday and Splunk and Informatica and Infor and Acquia and Datadog and Databricks, just a much broader collection of software that when you're moving to the cloud, you can use really easily on top of our platform.

But the other interesting thing that we see is that companies are moving to the cloud with -- very often with different systems integrators than they've used on-premises for the last number of years. And a lot of that has to do with the fact that companies, when they're making this big a change, they want to work with an SI that's really deep in a cloud platform and has a lot of dedicated trained professionals on that platform. And I think a lot of the large SIs have had this dilemma because they've had these big businesses of doing outsourcing and where they build practices and hedge their bets across lots of different companies. And if you don't actually get deeper in one of these platforms, the companies that are taking this risk of moving get a little bit nervous.

So there's some big SIs that I think have made that shift pretty well and done so successfully. These are companies like Deloitte and Accenture. But what we see is that a lot of the heavy lifting of moving enterprises to the cloud is being done by SIs that have either pivoted their model quicker and realized what the future was. And these are companies like Slalom and Rackspace, or born in the cloud SIs who don't have to worry about cannibalizing their existing business and who are very happy to pick up the small pilot projects for all of you that don't pay very much. They don't seem like they're worth it. But everybody knows, you can't move unless you get pilots done successfully. So they're willing to bet on the future. And these are companies like Onica and Clearscale in the U.S. and Cloudreach in the U.K. and AllCloud in Europe and Versent in Asia Pacific who are doing a lot of the heavy lifting. So lots of different choices in partners as people are moving to the cloud as well.

So when you think about transformation, there are some transformations that you have to do to yourself: deciding you're going to make that move, energizing your company to make that move, equipping your developers with the broadest possible capabilities that you can make that move quickly and successfully and efficiently and then making the hard decisions about what you're going to take with you in that new era versus what you're going to leave behind. But there are also transformations that are really about meeting new environments or new technical challenges. And I think if you look at the age we're in with respect to how much data that we are trying to store and process and analyze, it's like nothing we have ever seen before. This is a very different era than what's existed in the past.

You're not in the world where you're storing mostly gigabytes of data and sometimes terabytes. We're in a world where you're consuming petabytes of data and sometimes exabytes. And there's lots of reasons for it, some of it has to do with the borderless Internet and all the applications that can reach everywhere in a much broader way than was possible before. But a lot of it has to do with the cloud and how much more cost-effective and accessible compute and storage are and how much faster you can get work done. And the companies who thrive in this new era of this much data are not using the same old tools they've used forever. They are adapting to meet the new technology challenge and the tools that are required to do so. And so one of the challenges that you find for companies is that they have all this data they've accumulated over a long period of time. And unfortunately, they live in these data silos all over the place, which makes doing analytics and machine learning painful, expensive and slow. And it's why companies are so excited about being able to build data lakes to bring all that data together to make analytics and machine learning much easier.

And if you look at what people are building data lakes on top of, there's nothing that comes close to approaching how many data lakes have been built on Amazon S3, our objects store. And there are a few reasons why people are choosing S3 as their data lake base, first is it just turns out to be more reliable and scalable and available than anything else. And a lot of that has to do with our multi-availabilities and our multi-AZ architecture where we will take your data and we'll store it in multiple data centers, typically 3. And those data centers will be a few miles apart, no more than 100 so that you have the right low latency to operate for your application. If you compare that to what other providers do, they either mostly have regions where there's only one data center or if they have a multi-AZ capability, the AZs live in the same building or they're right next door to each other.

So if there's a problem in that building or on that street, that's the end of your availability and durability story. So very different on availability and scalability.

Second, S3 is the most secure data storage that you have. And there's lots of examples that I can give. But when we talk about this era of giant data like we are now, you really want to have granularity down to the actual object level. And so in S3, it's the only object store that allows you to block public access at the bucket and account level. It's the only object store that gives you inventory reports on all your objects so you can answer questions like, "Are all my objects encrypted?" It's the only object store that allows you to analyze all the access permissions on all your objects and buckets with the feature we launched yesterday called IAM Access Analyzer.

So most available, most secure. It also is the object store and data lake that gives you the most ways to get your data into it. And that turns out to matter a lot because you have data coming from everywhere. So you can get it in through the wire. You can get it in through our Direct Connect, through our backbone, through streaming and through IoT, a storage gateway, through SFTP if that's what you want to do, through the Snowball appliances, even through a 45-foot container with Snowmobile. Almost every magical way you can imagine getting your data from wherever it is into your data lake, you can do in S3 significantly more ways than you'll find elsewhere.

The other thing is that AWS actually gives you the ability to automate more of your actions which, again, if you think about the era we're in, a very large data matters. And so you can look at S3 Intelligent-Tiering, which is a storage class unlike anything you'll find anywhere else, where we assess using a machine learning-powered algorithm which objects are hot and which are being less frequently accessed and we move it to warmer or colder storage and adjust your price accordingly. Or if you're operating on hundreds of petabytes of information, sometimes exabytes with our customers, you don't want to have to take operations across every single one of those objects. You want something like S3 Batch Operations that lets you take actions across all those objects, again, something you won't find elsewhere. Then S3 has the lowest cost options with S3 Glacier Deep Archive. You won't find anything more (cost effective). That's cheaper than tape, by the way.

So those are the things that you find in S3. You won't find those capabilities elsewhere, which is why so many people are choosing S3 as their data lake. But if you think about this world where you now have data lakes. And we have a lot of customers who have very large data lakes now, it's actually pretty challenging. You have all of these applications, all these places you're setting in data; all this data being aggregated, normalized and indexed and reformatted. You have all these applications that are trying to access this data lake and lots of people. And for the person who has the job of having to decide which applications should have access to which data, it's very complicated. It's stressful. And we have a capability in Identity and Access Management, or IAM, that gives you all kinds of flexibility in what you do there.

And so customers have used that flexibility. But you have all these different access policies and objects. You have to configure who needs what. And people often make mistakes. And customers have asked us if we would help them find a way to simplify that, especially in this age of very large data. So I'm excited to announce the launch today of Amazon S3 Access Points, which radically simplifies managing access at scale for applications using shared data lake services.

So Access Points give you a customized path into a bucket with a unique host name and access policy that enforces the specific permissions that you've set up. And they're very flexible. You can also have access points that lock down access just for VPC access. So you can lock down your network. Then now if you have to think about deciding who should have access to what data and only that amount of data, it makes it so much easier. You don't have to layer on thousands on top of with one bucket. You can assign a different access policy to each application. It scales much better. You're going to make way fewer mistakes, much, much easier. This is a very unique and helpful feature when you're building a broad data lake.

Now as companies are moving to these data lakes. And we have tried for a couple years now and we'll continue to try to make it easier and easier for you to manage these very large-scale data lakes. And S3 Access Points is an example of that lake formation which we announced last year in the Keynote which makes it much faster to build a data lake from scratch is another example of that. But as you have those things, at the end of the day, if you don't have the right analytic services, it doesn't matter.

And so you want the broadest capabilities in the analytics space. And that's what AWS has provided for you over the last number of years. Nobody has the selection of analytics services we have. If it turns out you want to do ad hoc query of unstructured data, you can use Athena. If you want to process vast amounts of unstructured data using dynamic clusters of distributed frameworks like Spark and Hadoop and Presto and Pig and Hive and YARN, you use EMR. If you want to do superfast queries on structured data like a data warehouse, you use Redshift. If you want to do analytics on real-time streaming data, you use Kinesis. If you want to do BI with beautiful visualizations and great embedded and ML capabilities like you won't find elsewhere, you use QuickSight. And so very broad array of analytics services. But I want to go back to Redshift.

So when we launched Redshift in 2012, it really changed the data warehousing space. It was the first data warehouse built from the ground up for the cloud. And it really changed the pricing equation where it was less than $1,000 per terabyte. And it was the fastest-growing service in AWS for the first three years until Aurora was launched. It's continued to be a very fast-growing service for us. And we have tens of thousands of customers who are using Redshift. It is the most broadly used data warehouse in the cloud. And these are companies like Electronic Arts and Aetna and McDonald's and Yelp and Dow Jones and Pfizer, Intuit, Liberty Mutual. And we have more and more customers gravitating to Redshift in significant part because we're continuing to iterate at a fast clip. So if you look at over the last year alone, the Redshift team has added over 100 features. And I won't share all of them with you. But I'll touch on a few that I think have really mattered to our customers.

The first is concurrency scaling, which we launched about a year ago, which automatically adds and removes capacity based on unpredictable demand. And so it's incredible how many of our Redshift customers are already using concurrency scaling. And because we give an hour a day of free usage of concurrency scaling, about 97% of you who are using it are using it for free.

Second, just a few days ago, we released materialized views, which takes your most frequent queries and precomputes them and caches the aggregations and the filters and even joins between tables, which makes your queries go much faster. And people are pretty excited about that.

The third thing that we've been working on now for a year or 2 is something that we called Spectrum but we now refer to as the lake house, which is really about being able to query not just the data that you have stored locally in nodes in Redshift. But also across your data lake in S3. And not surprisingly, as people start querying across both Redshift and S3, they also want to be able to query across their operational databases where a lot of important data sets live. And so today, we just released something called Federated Query, which now lets you query across Redshift, S3 and our relational database services, including Aurora Postgres.

Then when you're doing this querying across all these data stores and getting these aggregated data sets in Redshift, customers want to move those data sets back to the data lake because they want to let all the other analytics services and machine learning services to be able to use those as well. And that actually turns out to be difficult and a pain in the butt to actually do. You have to do all kinds of work. And so we've made that easy for you with a new query called Data Lake Export that we're releasing today as well for Redshift. So Redshift is continuing to iterate at a very fast clip based on a lot of what you're telling us matters to you.

So when you step back and you look at Redshift, it's the most broadly used data warehouse in the cloud. It's 2x faster than anything else out there if you don't fudge the benchmarks. It's 75% less expensive than anything out there. And yet, what we would argue is that as you look at the age of data that we're in today and if you fast forward even just a couple few years how exponentially the amount of data that people are storing and trying to process and analyze is going to increase, you have to be looking around the corner and adjusting and evolving to allow you to do what you want.

And so we've thought a lot about this in this space. And one of the first things that you all have told us is, "Hi. look. I really want to be able to scale my storage and compute separately in Redshift. If I have -- Redshift has these instances to have both storage and compute contain in them, if it turns out that I have a workload and needs more storage, I have to provision other instance even if I don't need a compute, I want to scale those separately." Which seems like a pretty reasonable request. And so I'm excited to share with you our way to allow you to scale storage and compute separately with our new Redshift RA3 instances with managed storage.

And so what these RA3 instances have is very fast, big SSDs in the local instance. And if it turns out that you have a workload that exceeds the SSDs -- the amount of storage in the SSDs in the local instance, we've built technology that intelligently and automatically will move the less frequently accessed data to S3. But then what we did is because we have Nitro, like I was talking about earlier, we built unique instances that have very fast bandwidth. So that if you actually need some of those data from S3 for a query, it moves much faster than if you just had to leave it there without that high-speed bandwidth instance. And so with RA3s, you get to separate your storage from your compute. If it turns out, by the way, on your local SSDs that you're not using all the SSD in the local SSD, you only pay for what you use. So a pretty significant enhancement for customers using Redshift.

At the same time, if you think about the prevailing way that people are thinking about separating storage from compute and letting people scale separately that way as well as how you're going to do this large-scale compute where you move the storage to a bunch of awaiting compute nodes, there are some issues with this that you got to think about. The first is think about how much data you're going to -- at the scale that we're at but then just fast forward a few years, think about how much data you're going to actually have to move over the network to get to the compute. And we have very, very large-scale networking capacity, multiple petabits per Availability Zone. We have 100 gigabits per second instances. We have a very large, scalable network.

But it's not that hard to look around the corner and realize that it's going to saturate the network at some point and it's going to slow down performance. And that's going to be a real bottleneck for you. And if you could get through that networking bottleneck, which I'm not sure you can but let's imagine that maybe you could, then you've got a second problem, which is if you look at hardware trends, what you'll notice is that the throughput in SSDs to and from them and between nodes in SSDs is scaling and growing at 6x faster rate than the ability for CPUs to process data and memory. So it means even if you get through the networking bottleneck, the CPUs won't be able to keep up with the storage. And that means that you're going to have a performance problem, unless you decide to provision more compute. But then you're adding cost and you're not separating the storage from the compute again.

So this led us to really think about what can we do? How do we need to evolve to allow you to have this better performance in this world that we're moving into? And so the team has spent the better part of a year or 2 thinking about this. I'm excited to announce for you AQUA, which is Advanced Query Accelerator for Redshift, which is an innovative way to do hardware-accelerated cache that lets you build up to have -- lets you have 10x better query performance than any other cloud data warehouse solution out there.

So here's what AQUA does. First of all, it totally flips the equation of moving the storage to the compute on its head, where we're moving the compute to the storage. And what we built with AQUA is a big, high-speed cache architecture on top of S3. And the cache can scale out in parallel to lots of different nodes. And in each of the nodes, we have AWS-designed processors to make things go much faster. So we've taken a Nitro chip, adapted it and innovated on top of it to allow you to speed up compression and encryption.

And so this makes your processing so much faster that you can actually do the compute on the raw data without having to move it. Then it also saves you a lot of work because in the past, there's all this work to actually build data movement pipelines and the precompute things. The movement of all that data from the storage to compute is a lot of muck. And so with AQUA, you get the double benefit of it being much faster and being able to do the compute on the raw storage and saving time and energy from not having to do the muck of moving the data. This will also give you 10x better query performance than you'll find anywhere else. It will work with your existing Redshift implementations. We're going to do all the work to make the migration simple and easy. And it will be available for you in mid-2020. So we're really excited about this.

And I think it's also a pretty good example of something that we see all the time, which is when you build something new, it's only new for a period of time. And when you build something that's shiny, it's shiny until something's shinier. But the great products and the great companies find a way to carefully listen to customers and relentlessly keep innovating on their behalf. And you've seen this not just in the Redshift space. But in compute, in storage, in database, analytics and machine learning from AWS in our first 13 years and you should expect it to continue.

Now when you think about the scale of data and some of the ways we're changing Redshift to allow you to manage it even looking forward a few years, Redshift is not the only analytics service that you have to think about this. You also have to think about it in areas like Elasticsearch.

And so many of you use Elasticsearch in this room. We have a managed Elasticsearch Service that we launched a few years ago, which is growing like a weed. It's incredible how much people are using this service. We have tens of thousands of customers, very large customers like Nike and Intuit and Airbnb and Hulu and Pinterest. And as these customers think about their use of Elasticsearch and how the amount of data is changing what they want, they also realize that there are challenges. There's this explosion of data. Because so many people are building with microservices now, the amount of log data that people want to use to monitor and assess their operational performance is just gigantic.

And one of the extra challenges in Elasticsearch is that the file format is optimized for search, not for storage size. So it's relatively inefficient. And we have a number of customers where if they want to actually store months of operational data, it's many hundreds of terabytes of data. And so what happens is it's expensive enough that customers don't do it. Most of our customers are storing just a few days, maybe a week's worth of operational data in Elasticsearch. And that's not what they want. There's a lot of reasons why you want to analyze your operational data, your log data over a longer period of time.

So this, again, is something that we've thought about. And we've tried to figure out if we could build a solution that will work for all of you. And I'm excited to announce the launch of UltraWarm, which is a new warm tier really on steroids for Amazon Elasticsearch Service.

So typical warm storage layers for Elasticsearch Services aren't used very pervasively because the performance is pretty laggy and the durability is not very good. And so we've taken a different approach in how we built UltraWarm. It's designed to be a warm tier on steroids with much better durability that's backed by S3. And there are several things that UltraWarm does that are a little bit different.

First of all, what we look at is very sophisticated technology and advanced placement techniques to look at the blocks of data, down to the blocks of data that are being frequently accessed or not frequently accessed. And the ones that are not being frequently accessed, we move to S3 so you can save money. And by the way, we think with UltraWarm, if you use it right, you'll save about 90% on your storage costs versus what you do in Elasticsearch today and 80% less expenses than any other warm tier you'll find out there for an Elasticsearch service in the cloud.

Then, again, leveraging Nitro and unique instances that we built with fast bandwidth, we allow you, if it turns out that you need to make a query that pulls data from S3, we have this very high bandwidth instance that makes that much faster so you get the snappy, interactive performance that you need and that you expect when you use Elasticsearch. So we're very excited about providing this for you today. It's easy to sign up for. The preview starts today. And we're excited to have you take a chance with it.

So if you think about how the amount of data is changing what you do and what you use, it's changing how you think about data lakes. It's changing how you think about the access policies to give lots of people access to your data lakes. It's changing the analytics services you're using. It's changing the price performance that you're going to expect and need to be able to process the amount of data that you want to in these analytics services.

But it's also impacting databases. And databases are not immune to this issue. You have all the same types of challenges that we talked about with analytics. And for a couple of decades, a lot of companies primarily use relational databases for every one of their workloads. And the day of customers doing that has come and gone. There is just too much data for it to make cost sense and complexity sense to do that anymore.

And so what's happened is that this has demanded customers to ask for and really demand purpose-built databases. And so if you actually are a company like Lyft and you have millions of drivers and geolocation coordinates, you don't want a relational database that's too expensive and too complicated, not performing enough. You want a very fast, high throughput, low latency, key value store, which is why we built DynamoDB. Or if you have a workload that requires sub-microsecond latency, you want an in-memory database, which is why we built ElastiCache for Redis and Memcached.

And if you want to connect data in multiple big databases of data, you want a graph database, which is why we built Neptune. And if you're doing a lot of IoT like many of our customers are, we're actually -- what you're trying to measure is the change over time. You want a database that's anchored on variable time, time series. And so that's why we built and announced Timestream. If you run a supply chain and you want to have some type of transparent and mutable, cryptographically verifiable ledger, you want a ledger database. That's why we built QLDB. Or if you're a company that does a lot of work with JSON documents, you want a document database, which is why we built DocumentDB with MongoDB compatible. This is a set of purpose-built databases that have come from things that we've listened to that developers care about to optimize their customer experience.

And so a lot of our developers look at this list of purpose-built database and say, "This is pretty awesome. You have a selection unlike others. But there's one obvious missing one that I don't understand why you don't help us with." And that obvious one is Cassandra. And people say, "Well look, we manage Cassandra on-premises." And what I'd tell you what it's like to manage Cassandra on-premises, it's kind of the same story on all these things and why we're trying to move into the cloud. It's hard to manage the hardware. It's hard to manage the software. It turns out it's really difficult to scale up and down. So we all scale up for the peak. So we're sitting on a lot of wasted costs. The rollback features in Cassandra are pretty clunky. So people are often operating on old versions of Cassandra, which is dangerous for obvious reasons in security.

And so customers said, "Why don't you do something about it?" A lot of our customers, when they get to very large scale at Cassandra, they tend to move to DynamoDB. And this is what companies like Nike and Samsung have done. But understandably, companies have said, "I don't want to have to move. I want to be able to use the Cassandra interface if I want to as I scale." And so I'm excited to announce today the preview of Amazon Managed Cassandra Service.

And so with this new Managed Cassandra Service, it's compatible with the 3.11 release, no clusters to manage, single-digit-millisecond latency with all your workloads. You can choose to provision a certain amount that you think you need, you know must be there at a certain point, or you can just choose to provision on demand and pay on demand. It uses all the same Cassandra tools and drivers. So it will be easy to migrate your Cassandra workloads from on-premises to the cloud. Then we've integrated it across all of our various AWS platform capabilities so you can use it as part of the platform. So I'm excited to give Cassandra users this capability today.

So when you think about this collection of purpose-built databases, you won't find this collection anywhere else. And when you have a company that says, "I don't know. You don't need that many databases. I have a relational database. And it can take care of all this for you." You should nod and say, yes. Then some companies say, "No. I have a non-relational database. And it does key value really well and it does document really well. And it does graph really well. And does it time series really well." You should listen, you should be polite. And you should be very skeptical. Swiss Army Knives are hardly ever the best solution for anything other than the most simple task. If you want the right tool for the right job that gives you differentiated performance, productivity and customer experience, you want the right purpose-built database for that job. And we have a very strong belief inside of AWS that there is not one tool to rule the world, that you should have the right tool for the right job to make you spend less money and be more productive and change the customer experience.

Now as I mentioned earlier, there are some transformations where you have to transform yourself. And there are some transformations where you have to adjust to the changing environment, like what we're seeing with a huge amount of data or some changing technology. And some of those transformations are relatively understood walks. And some of those are a little bit further and less well understood journeys. This has been the case in machine learning for the first several years, which is that developers and data scientists and companies are so passionate and so excited about getting value out of their data with machine learning that they've been willing to deal with clunky tools and walk those 500 miles to get there. However, as many of you know, most people aren't willing to walk 500 miles. And so a lot of what we've tried to do in machine learning over the last few years in AWS. And continues to be our mission, is to allow you to get what you want to get done and get value from your data using machine learning much easier than ever before. And we have a lot of machine learning that's happening in AWS. And it's grown very substantially. I think sometimes, people forget that machine learning isn't just something called machine learning. It's -- you got to start with the right, highly reliable and available and scalable data store. And then you need the right access control and the right security on top of that and the right analytics. And then the right machine learning capabilities. And there's nobody who has that collection of capabilities for machine learning like AWS, which is why we have tens of thousands of customers that are using us for machine learning and twice as much more machine learning in AWS than you'll find anywhere else. And lots of companies that you know, Intuit and FICO and GE and Change Healthcare and Guardian, Volkswagen and the NFL and NASCAR, Formula One and Expedia, in Ford, Dow Jones, just a very broad array of companies that are leveraging AWS for machine learning.

And when we look at the future of machine learning. One of the areas that we have become increasingly excited about, that we think can completely change the world and change outcomes for everybody in and out of this room is how machine learning can help health care. And one of the leading companies of health care in the world is Cerner. And I'm going to welcome to the stage the CEO of Cerner, Brent Shafer, to share how they're transforming health care with AWS.

David Brent Shafer {BIO 19841210 <GO>}

Thank you, Andrew. And thanks to AWS for the opportunity to be here with you today and share our story. At our core, Cerner is a health care technology company. And we developed software that powers health care delivery throughout a patient's lifetime across multiple venues and providers of care. And we do that while helping health systems efficiently manage complex billing and revenue cycles.

Now we're a global company, managing the data of nearly 250 million people around the world. And every day, close to 3 million health care professionals in more than 30 countries access our secure system. And at 23 petabytes, Cerner has both one of the largest collections of personal health information in the world and a tremendous opportunity to transform the well-being of the world's population.

Our transformation is something we're accustomed to. For 40 years, we've ushered in health care's digital age, by moving medical data from paper charts to Manila folders to electronic health records. And this process is nearly complete. And it's providing a more organized view of patient medical history. And it's also improving communication among care teams and, overall, the quality of care.

But as you know, health care is a journey of continuous improvement. And we have a lot more to do. Globally, if you think about it, there's more than $7 trillion spent on health care each year. And the United States alone accounts for roughly half of that. So -- and recently, the American Medical Association estimated that up to 1/4 of U.S.

spending on health care, that's nearly $1 trillion, is wasted. And that waste comes in many forms. It comes in variation of care delivery. It comes sometimes in overtreating patients. And we know for sure, we have data gaps that lead to challenges as patients age and receive care from different doctors and different facilities spread out across the country. So the volume of data in care delivery requires new tools.

So let's start by thinking bigger. What Cerner is really doing is making the world's health care data actionable. For example, we made paper data more accessible by going digital. And now we're working on reducing variation on how providers actually deliver the care. But what we and the rest of health care really haven't done well yet is learning, predicting and preventing problems by leveraging the power of the data.

So who has done this before? Who knows how to leverage the power of data really well? Amazon Web Services. So we're delighted to be here. And in July of this year, we announced an expansion of our AWS collaboration to help us drive our strategic priorities. And those are around migration, modernization and innovation. So we're migrating our privately hosted platforms to AWS. And we're doing that with our joint pledge for the responsible and ethical treatment of health care data. And we want to modernize the way we deliver our solutions by enabling Software-as-a-Service and Data Science at Scale. And it's really by using infrastructure and machine learning services that will help us do that.

And we're innovating by leveraging these services into new solutions for the marketplace. Our goal, overall, is to improve patient outcomes. It's to reduce administrative and operational complexity. And it's to predict and prevent health issues as early as possible. So let's look at 2 examples. As you may know, you hear about it a lot in the press, unnecessary follow-up visits to a hospital are a huge contributor to waste in the health care system. So we call these readmissions. And a readmission is basically a redo. So think about taking your car into the local mechanic to have something fixed. And within 30 days, you're back in the mechanic shop, having it done again, basically the same work. Very expensive and very frustrating. Ideally, that second time of that readmission should not happen. So in health care, readmission costs are -- actually, they're higher than the initial visit for 2/3 of the most common diagnosis. And if we could predict that readmission and do something to help the clinician to modify the treatment plan, we can possibly prevent it in the first place. And of course, that's much better for the patient, it's better for the caregiver, it's better for the family and it saves on the cost.

So one of the largest health care providers in the United States asked us to help them predict patients who are at risk for being readmitted. So we leveraged five years of AWS data that we had aggregated together with them over -- our data aggregated with AWS. And we use machine learning services to build, train and deploy their predictive model. Now with that prediction, the caregiver can now change their approach before discharging the patient. And that's really the key to preventing second episodes of care. And these are often very significant for patients that have traumatic brain and spinal cord injuries, stroke and other neurological conditions and many more.

So this model is creating the knowledge, skills and capabilities of both the health care system and Cerner. And the collaboration was really made possible through the cloud infrastructure provided by AWS. And as a result, the health care system reported its lowest readmission rate in more than a decade while simultaneously increasing its discharged community rate. All that means basically fewer patients were being readmitted. And more patients were returning to their homes, which is, of course, is where they would prefer to be. So as we leverage more of the infrastructure and machine learning services by AWS, we expect to see more successes like this.

Another example of the work we're doing with AWS is really focused on returning the joy of practicing medicine. And if you read the popular press, we hear about it. One of the things you'll hear is that in the United States, about 40% of physicians report that they feel depressed or burned out by the stress of the role. And part of that is certainly that they're spending more than half their day often documenting information, just doing data entry. So this has really become a burden. So yes, we've digitized health care. But at the same time, the documentation requirements have gone way up.

So what if we could reduce or eliminate data entry for the physician? I mean think what an incredible opportunity that is. So Cerner is developing a virtual scribe application that captures doctor-patient interaction using speech recognition. What it does is suggest allergies, medications, medical problems. And it integrates that information directly into the physician's workflow. And we're using AWS Transcribe Medical for the speech recognition that powers this innovation. And I'm confident with AWS' clear leadership in voice and speech. And our expertise in health care workflows, will give doctors more time to spend with patients. And that's what they really want.

So we're excited about how this collaboration helps us move closer to Cerner's vision. Our vision is really a seamless and connected word -- world where everyone thrives, created by breakthrough innovation that shapes the future of health care. Because we can all appreciate that health care is far too important to stay the same. Thanks for your time this morning. I hope you have a great day. Thank you very much.

Andrew R. Jassy {BIO 15111610 <GO>}

Thank you, Brent. It is really an honor for us to be partnered with Cerner the way we are and be a small part of helping them change health care outcomes for all of you. So really appreciate the partnership.

So I thought what I would do is spend a little bit of time sharing how our view of machine learning continues to evolve. And we continue to believe that there are 3 macro layers of the stack of machine learning. The bottom layer is for expert machine learning practitioners who are very comfortable with the framework level. And this group deals with the 3 primary frameworks: TensorFlow, PyTorch and MXNet. And TensorFlow has the most residents and the largest community today, continues to do so. And we have a lot of TensorFlow experience. If you look at the amount of TensorFlow that runs in the cloud, 85% of the TensorFlow in the cloud runs on AWS. So we have a lot of customers running it. And we have a team that does nothing but work on optimizing TensorFlow performance on AWS. And they have done a lot of innovation. They have built the fastest-growing TensorFlow you'll find anywhere. And they've done things where they've really changed the scale and efficiency to achieve close to linear scalability across hundreds of GPUs. And they've done this by inventing, innovating on a way in which TensorFlow shares model parameters between multiple instances, making the sharing faster and more efficient between these instances. And so we have the separable team just on TensorFlow to optimize the performance. But this is where I think we also have a pretty big difference from what others do. Most other cloud providers try to funnel everybody through TensorFlow. Some have started to support a little bit some of the other frameworks, really the vanilla versions out of the box, not tune, leaves all the work for you. And it's a little bit of a self-fulfilling prophecy on why everybody is using TensorFlow. But one of the things that we have realized as we've done research. And we've talked to developers and to data scientists, is that 90% of data scientists use multiple frameworks. And that's because algorithms are being invented all the time by people all over the place in every type of framework. And people don't want to have to take the time to port that algorithm back to TensorFlow. And so we support all 3 of the major frameworks equally well.

And we have dedicated teams that not just work on TensorFlow. But the team that works just on PyTorch and just on MXNet. And it yields different results. Let me give you an example. If you look here, this is a common computer vision algorithm called Mask R-CNN. And the previously fastest time to run this was from a company in Mountain View that did it in 35 minutes using hardware that's not available to any of you. It's in some kind of private data. And if you look at our team that worked on TensorFlow and AWS, they -- because of the innovation that they've done that I mentioned, they're able to get it done 20% faster in 28 minutes using P3 instances, which are available to you. But because we actually care about and support all of the major frameworks, it's not just TensorFlow. Our PyTorch team and our MXNet team also optimize how this algorithm ran on it. And actually did it in 22% faster time in 27 minutes. And so we are always going to give you all of the major tools that you need to do your job. We're not going to make decisions for you of what you must use. We will give you the right tool for the right job.

Now there aren't that many expert machine learning practitioners in the world. Most of them exist, kind of hang out at the tech companies. And so if you really want machine learning to be as expansive as we all believe it can and should be, you have to make it more accessible to everyday developers and data scientists. And that's why we built SageMaker, which we launched a couple of years ago, which is really a C-level change and the ease with which developers and data scientists can build, train, tune and deploy machine learning models.

And it's incredible how many customers have started using SageMaker already. We have tens of thousands of customers. And we have thousands who are standardizing on top of SageMaker: Avis and Bristol-Myers Squibb and Chick-fil-A and CondÃ© Nast, Dow Jones and GE and Hurst and Liberty Mutual, Panasonic and Siemens, very broad group. And this SageMaker team has gotten a lot of great input from all of our developers. Thank you. Keep it coming, please. And they have been hard at work. They have launched over 50 features just in the last year. And just a few I'll touch on.

About a year ago, we launched Ground Truth, which makes it much easier for you to label your data. We gave you a marketplace, which gives you hundreds of algorithms from others that you can use in your machine learning. We were the first to build reinforcement learning and -- into a service like SageMaker, which people have been using in conjunction with DeepRacer for a year. By the way, the DeepRacer championship right before Werner's keynote on Thursday. You've always been able to do One-Click Training with SageMaker. But now you can do it on spot. If you will, it may take a little bit longer for your training and save up to 90%. Then we built something called Neo, which lets you train once. And then compile and practically every imaginable place on the edge. And so these are the types of capabilities that makes SageMaker so much easier to use and so popular. And while customers have said, "You have made it so much easier to do machine learning than it was before. And you've made every step easier, it's still true that all of the work in between those steps, in giving visibility and figuring out what's going right or wrong, is still a lot harder than we wished. And the same -- you could say, it's always true in software development, which is why we have the integrated development environments or IDEs. The problem is that there's never really been an end-to-end IDE in machine learning until now.

I'm excited to announce SageMaker Studio, which is the first fully integrated development environment for machine learning. So SageMaker Studio is a web-based IDE which allows you to store and collect all the things you need, whether it's code or notebooks or data sets, or settings or project holders, all in one place, one pane of glass. And it makes it much easier to actually manage all of those pieces in building a model. And so I thought I would share with you some of the things that are part of SageMaker that we're making available today.

So first are notebooks. And I think a lot of people know what notebooks are, a machine learning. But they're a place that people use to build machine learning workflows. And they contain sections of code and documentation and visualizations and results. And in SageMaker, notebooks are paired with compute. And if it turns out that you need more compute or less compute, you actually have to go and spin up another instance. And then you have to do all the work to transfer the contents from the first notebook to the new notebook. And it's just a little bit tedious. And customers ask us if we'll make that easier. So I'm excited to announce the launch of SageMaker Notebooks, which is one-click notebooks with Elastic Compute.

And so now you can just spin up a notebook with a click, it happens in seconds. If it turns out that you need more compute than you thought in this notebook, you just tell us, the CPU, that you want with that notebook and we manage that compute for you. And we do all the heavy lifting of transferring all the contents from that first notebook to the second notebook. And then shutting down that first notebook, if that's what you want. So a much easier way to manage notebooks. And people said, "Well that's great. Notebooks are such an important part of doing machine learning.

But let me tell you about another problem we have, which is when you're doing machine learning, you're trying all kinds of experiments and you're iterating like crazy across lots of different parameters and dimensions. And as you iterate a lot, it creates all these artifacts and they live all over the place. I can't find them and I can't share them. Please, make this easier." So I'm excited to announce the launch of SageMaker Experiments, which is a way to capture, organize and search every step of building, training and tuning your models automatically.

And so with SageMaker Experiments, it allows you to capture all the input variables, all the parameters, the configuration, the results automatically. And it saves in what SageMaker calls an experiment, you can have multiple experiments in a project. And now you can not only browse your active experiments and see them in real time. But you can also search for older experiments: by name, by input parameters, by data set use of the algorithm or even the results. And so it is a much, much easier way to find, search for, collect and share your experiments as you're building a model. So much easier way to manage notebooks, much easier way to manage experiments. People say, "Well how can you make training easier?" And training is actually quite different for lots of different reasons, not the least of which is that you're trying to work across dozens and dozens of parameters. And a lot of times, you don't really know which dimensions are really impacting the model.

Looking at a trained model is a little bit like looking at a compiled binary to understand how an application works. It's just totally opaque. It's like gibberish to the naked eye. And people want to have better idea of what's driving their model so they can adjust it and fix it. And so they can explain it. And so I'm excited to announce SageMaker Debugger, which allows developers to debug and profile their model training to improve the accuracy of their machine learning models.

And so with Debugger, it's on by default. We've done a bunch of work with all 3 major frameworks in SageMaker. So it's automatically setting to you the metrics that you want to monitor and to see what's actually happening. Then we have this capability in Debugger called feature prioritization. And what it does is it puts a spotlight on the actual dimensions or features that are having impact on the model. And so this -- it does 3 very useful things. It means, first, you actually know what's driving the model, which is hugely helpful as you're training a model. Second thing it does is it turns out if you have an underperforming neural network model, you might want to know which dimensions it's leaving out, that could actually help you understand why you're getting predictions that don't match what you think should be the case. Then also, if you have models that feature prioritization shows that are overly reliant in just a few number of dimensions, you might have bias in your model that you want to change. So very useful to help you train, understand what matters and also be able to interpret your model. So easier to manage notebooks, experiments, the debugging and profiling capability.

And people say, "How can you help me when I have models that have been working for a long period of time. And all of a sudden, it looks like the predictions aren't relevant anymore?" And this happens sometimes, like if you take an overly simplified example, let's say that you built the model in 2016, the estimated housing prices.

Well the model worked well in '16. And it worked well in '17 primarily because the conditions were the same. But then in 2018, as interest rates changed and housing prices went up, the model stopped making accurate predictions. And this is a concept in machine learning that people call concept drift. And it's actually, there -- if you know there's concept drift, you can actually make changes to the model.

But it turns out that the overwhelming majority of models are way more complicated than the simple example I just gave. And it's really hard to find the concept drift. It's all kinds of data wrangling, where you have to look at what was the model that the -- what was the data the model was trained on? What's the model that we're making predictions on now? And how are they different? And how are they changed? And when? It's just really complicated. So we're excited to help solve that for you today with SageMaker Model Monitor, which is a way to detect concept drift by monitoring models deployed to production automatically.

And so with concept drift, what we'll do is we create a set of baseline statistics on the data in which you train the model. And then we actually analyze all the predictions, compare it to the data used to create the model. And then we give you a way to visualize where there appears to be concept drift, which you can see in SageMaker Studio. And you can take charge of that and figure out how to make adjustments when you've got the situation of concept drift. So all of these things I just mentioned, notebooks, experiments, training with debugging and profile. And concept drift, are assuming that you are building models.

But we know there are a whole bunch of data sets that are very useful. But that people just don't have the time or the wherewithal or the capabilities to train a model for those data sets. A simple example would be, let's say, you had an operational database data set where it was all your sales leads. And then the leads that actually ended up becoming real sales. If you can actually build a simple model that predicted what the variables are, where leads convert to sales, you would spend your scarce resource and follow-up on those particular opportunities.

And so that's where there's promise that we've talked about in the past of having something that people call AutoML, or automatic machine learning models, has been. And there have been a couple of problems, though, if you look at it with these AutoML models, that people have tried to roll out. The first is that they build this okay, simple model initially that are total black boxes. So if it turns out that you want to improve a mediocre model or you just want to evolve it because it's something that matters to you, that can matter for the business, you have no idea how the model is built. There's nothing you can do about it. Or if it turns out that you want to make tradeoffs, maybe in some cases, you may not take the absolute best accuracy. You may trade a little accuracy for something like faster latency and prediction, given the nature of your application. You're out of luck. You have just this one simple black box model. And so customers have said, "We want AutoML. But we want more visibility."

So I'm happy to announce SageMaker Autopilot, which is AutoML with full control and visibility. And so with AutoML, here's what happens. You send us your CSV file with the data that you want to model for or you can just point to the S3 location. And Autopilot does all the transformation of the model to put into format so we could do machine learning. It selects the right algorithm. Then it trains 50 unique models with little bit different configurations of the various variables because you don't know which ones are going to lead to the highest accuracy.

By the way, even if you know how to build machine learning models, having to train 50 models takes quite a bit of time. So this is very useful. Then what we do is we give you in SageMaker Studio a model leader board, where you can see all 50 models ranked in order of accuracy. And we give you a notebook underneath every single one of these models. So when you open the notebook, it has all the recipe of that particular model. You can tell how it was built, you can tell all the configuration, you can tell the parameters, you can tell the algorithm. It gives you the whole recipe so that if you then want to take that model and evolve it and make it into something that really changes your business over a long period of time, you can. Then also, you can look at that model leaderboard with those 50 algorithms and you may look at the difference between algorithm 1 and algorithm 2. The difference in accuracy is tiny. But the difference in latency is significant that makes your application make predictions much quicker. You may choose to take that one. And so with Autopilot, it allows you to do AutoML in a way that not only lets you create a model automatically but gives you full visibility and control to be able to evolve that model and make trade-offs yourself.

So when you look at SageMaker, before today, SageMaker had become a C-level change in the ability to build, train, tune and deploy machine learning models. With SageMaker Studio, it's a giant leap forward, the first IDE for machine learning. And it's going to make it even easier for everyday developers and data scientists to build machine learning models. And to show you how the whole thing comes together, it's my pleasure to welcome to the stage as I do every year, the inevitable, Dr. Matt Wood.

Matt Wood {BIO 18000850 <GO>}

Thank you, Andy. Good morning, everybody. SageMaker Studio pulls together for the first time dozens of machine learning tools into a single pane of glass, which contains all of the tools that you need to build, train, tune and deploy your machine learning models. And our aim here is that we want to be able to provide machine learning and put it in the hands of even more developers and data scientists than we've done in the past. I'm going to give you a walk-through, a guided tour of SageMaker today using a very simple example. We're going to build a machine learning model which predicts house prices. We're going to do this using a couple of simple parameters, things like the number of bedrooms and the number of bathrooms for each individual house along with things like the mortgage rate and the price. And we're going to go ahead and collect this.

So machine learning learns by examples. And so the more examples you have, the better. We can collect all of this sales information from across the U.S. Then we simply take it, wrap it up into a CSV file and drop it into S3. Now we can move into SageMaker Studio. And in just a few clicks from inside the IDE, we can launch a new autopilot job. We simply just tell SageMaker where the model, where data is. What SageMaker Autopilot does at this point is it starts to spin up multiple different models, each with a different set of algorithms, data sets and parameters. The dirty secret of machine learning is that you don't just train a single model, you train dozens and pick the best one.

So here, SageMaker is using all of the information that it has to automatically pick and train the algorithm and the parameters to train multiple different models. And it does this iteratively. We're just showing 4 here. But in the next iteration, it picks the best features using machine learning under the hood in order to seed the next iteration and the next iteration. And over time, SageMaker Autopilot starts to home in on the best set of algorithms, data features and parameters to provide the best possible model. And it provides a ranked leader board of candidate models, here ranked by accuracy.

Now in almost all cases, you're going to want to choose just the best-performing model. So the one with the best accuracy. But because SageMaker Studio is using the debugger under the hood, it's providing profiling information and Autopilot is automatically generating the notebooks, you can dive into any one of these individual models and get a closer look. You can look at exactly how SageMaker Autopilot pull together all that data. You can look at the exact parameters and the exact algorithms that we use to generate it. And because we're using the debugger under the hood by default, you can start to inspect the features and their prioritization inside those models. So this allows you to, with new levels of visibility and insight, pick the best model which not only has the right accuracy but also meets other expectations such as whether you're treating the data correctly or whether your model potentially contains any bias. At this point, when you selected the model that you want to deploy, you can do that with just a single click inside SageMaker Studio. It gets deployed. And you can turn on the SageMaker Model Monitor.

And what we're doing here is we step by step start to compare the data which is used to make predictions with the baseline data used to train the model. And at any point, if SageMaker Model Monitor starts to detect any statistical deviations, it will give you an alert. And that's a really good indicator that you want to be able to go back, look at your model again and potentially retrain the data.

So here, we can see that the mean of the mortgage rate data has started to change. This probably has something to do with interest rates. So at this point, we can jump back into Studio, we can either pull up our notebooks. We can start from scratch using Autopilot and train another set of models, probably with new fresh data. Then we can go through the process again. We can select our best-performing model. We can deploy it into production, that goes into a fully managed elastic Multi-AZ environment and model monitoring will continue until it detects any deviation going forward.

So for the first time, SageMaker Studio starts to pull together the tools that developers are used to using with traditional software, debuggers, profilers, automation, management into a single pane of glass which can be used to build, train, deploy and manage our machine learning models in a way which is way easier and way more accessible for even more developers and even more data scientists.

And with that, I'll hand it back to Andy. Thanks.

Andrew R. Jassy {BIO 15111610 <GO>}

Thank you, Dr. Wood. I always love Dr. Wood's presentations. I appreciate it.

Okay. So we talked about the bottom layer of the stack for expert machine learning practitioners and the middle layer of the stack with SageMaker now SageMaker Studio for everyday developers and data scientists. And now there's the top layer of the stack, which we call AI Services because these services most closely mimic human cognition. And we have a broad array of services there. For vision, we have this thing called Rekognition, which is things like here's an object, tell me what's in it. Here's a video tell me what's happening in the video. We have for speech, we have text to speech with Polly. We have the ability to transcribe audio with Transcribe. And on the tech side, people want that transcribed text translated into multiple languages, which we do with Translate. We have an OCR++ service, which not only takes the data from pretty material but also in complicated formulas and tables, graphs, in Textract. Then we also have the ability to do natural language processing on top of all that data so that you don't have to do all the work yourself to read it and know what the heck is going on in there.

Then we have a number of services that are really borne out of things that we've done at scale at Amazon that you've asked us to expose as services for you. So we took all the natural language understanding and automatic speech recognition in Alexa and made that available to you at a service called Lex. Then last year at re:Invent, we gave you both the deep personalization and forecasting experience that we have at Amazon in services called Personalize and Forecast.

And so customers have asked us to think about other areas where we have really deep experience where we can help customers. And so one of the obvious ones is fraud. If you think about fraud, tens of billions of dollars around the world every year are lost to fraud. And people have fraudulence -- fraud detection services and systems. But they're pretty expensive. And they're pretty clunky and they don't use much machine learning. And they're hard to manage. And they have a lot of hard-coded rules that don't scale very well. And one of the things that we've realized in over 20 years of doing fraud detection in Amazon's Consumer business is that machine learning is unbelievably helpful. But for all the reasons we've talked about in walking 500 miles, it's hard for most companies to use machine learning in fraud detection. So we've thought about that. And we decided to announce today the launch of a new service, which is called Amazon Fraud Detector, which is a new machine learning service that does fraud management for you.

And so here's how it works. Transaction data, we take that data , along with the algorithms that we have built to detect fraud in Amazon's Consumer business. And we build a unique model for you. And what we also do is we have a set of data detectors that we've developed for our Consumer business in Amazon. And we take those data detectors, overlay them on top of the base model that we built for you and we create your own unique model that we exposed to you through a private endpoint API. And so what then happens is as you have new activities where there's sign-ups or online purchases, you use the API, we run them through the model. And we return a fraud score among other things. So that then you can take action, many of which you will automate based on the fraud score. So just a completely different way to manage fraud using machine learning that we're excited to give you today.

Now personalization, forecasting, fraud detection, automatic speech recognition, natural language understanding, these are all things that we have done at scale at Amazon that we have exposed as services. And I think one of the things that's different about AWS from other technology companies is I think that a lot of companies in the technology space like to work on things that they think are cool and where they're attractive to technology which, by the way, is understandable. We like working on cool things, too. But our priority and our focus is not on just working on cool technology or something that looks good in the press release. We are here, we spent all our time trying to build solutions that help you get your job done better and change your customer experience. So the question that dominates our conversations as we're thinking about what to work on and spend resources on next is what else can we build that can give value to you. And in places where we have done this over a long period of time, at Amazon, we're going to see if we can find a way to expose those as services so that you can get your job done better. And one of the areas, as we were racking our brains, that we thought we might be able to help further in is code. And most people in this room know this routine. You write code, you have to review the code, you have a mechanism to build and deploy the code, then after you do that, you measure it. And then you improve it and then you rinse and repeat. But the problem, of course, is if there's a problem with the code, these other steps don't really matter. You're going to have a bad customer experience. And that's why we all do code reviews and they're all manual code reviews. And there are a lot of organizations that don't have enough people do these code reviews or even their best people who do them miss things because they're moving fast, they have a lot of things going on. And so we thought about this issue and wondered if we might be able to provide some help. And so I'm excited to announce the launch of a new service today, which is called Amazon CodeGuru, which is a new -- it's a new machine learning service to automate code reviews and also to identify your most expensive lines of code.

And so as I mentioned, this service has 2 components. I'm going to start with the automatic code review. So what you do is you write your code, you commit it as you always do. And we'll support Github and CodeCommit to start with and we'll support other repositories over time. Then when you submit a code change and you do a pull request, you do that as normal. But you just add CodeGuru as one of the recipients to this pull request. Then what CodeGuru does is it goes through models and algorithms that we've built from millions of code reviews that we've done in Amazon over the last 20 years along with the training we've done in the 10,000 most popular open source projects. And it provides you an assessment of your code. And where we see there's a problem, we'll give you a human readable comment that will tell you what the issue is and point it out down to the line of code. And so what are some of the things that will let you detect?

So the first is AWS best practices, if you're missing a pagination or if you're not -- having error handling it right, or if you're using the APIs or the SDK features in a way that we think is suboptimal. When we shared CodeGuru privately with a few customers, just this first piece of adherence to AWS best practices was a game changer for them. But it also will identify concurrency issues. These are things like atomicity violations or using non thread-safe classes, which it turns out, these are pretty difficult to find. If you have incorrect handling like you're failing to release streams or database connections from memory or if you have unsanitary inputs that could lead to injection attacks or denial of service, CodeGuru will identify all these types of problems and issues for you in human readable comments to make it much faster and more reliable to do code reviews.

And the second part of what CodeGuru does. And this is a question that was borne out of a number of operational readiness reviews at Amazon, where we talk to different teams about their applications where there may be issues or not. And we're always asking ourselves, how can we find the most inefficient, unproductive cost expensive lines of code? And so that's the second piece of CodeGuru, which is really a machine learning powered profiler, simple to get started, you just configure it in the console. You install a small low profile agent on your application and then CodeGuru observes your application and every 5 minutes creates a profile. It will tell you things like latency and CPU utilization. And it helps you identify the most expensive lines of code that you have in your application to improve them. And this can make a big difference. We have used this at Amazon Now for a couple of years. We have 80,000 applications internally that are using the profiler part here of CodeGuru. And it's led to tens of millions of dollars of savings for us. So I'll give you a simple example, Prime Day, which is the largest e-commerce day in the world, we had our consumer payments team that was using CodeGuru for profiling. And they found most expensive line of code after most expensive line of code after most expensive line of code and made changes. And throughout the year, even as their application was growing very substantially because our Consumer business is growing quickly, they were able to improve their CPU utilization by 325% and saved 39% in costs from where they were before in just a year. So it makes a big difference. And we're very excited to give you CodeGuru today.

Now a year ago, in this keynote, one of the things that I talked about was that we have these 2 macro segments of developers and customers. And there are some builders who want access to all the low-level building blocks. So they have the flexibility to stitch the applications together however they see fit. Then there's another segment of builders who say, "I'm willing to give up some of that flexibility in exchange for getting 80% of the way there faster." They want a different level of abstraction. And the same is true in machine learning. We have loads and loads of customers that are using all those building blocks I talked about at that top layer of the stack. But we also have customers who say, "Look, I would actually like to have you stitch some of these together. So I don't have to do as much of the work."

So let me give you an example. If you look in our service called Amazon Connect, which is our call center in the cloud service, which is off to an unbelievably fast start, one of the fastest-growing services in the history of AWS with customers like Intuit and Capital One. And GE and Citigroup and John Hancock, Best Western and Johnson & Johnson and Hilton, is just off to a blazing start. And the reason people like this service so much is that it uses the same customer service technology that Amazon has used for several years. It's very scalable, easy to scale agents up and down. It's very cost effective. And it's really easy to get started to use. You don't have to be highly technical to use it. And it's the first call center in the cloud that's set up right from the get-go with the cloud and machine learning in mind. And actually, the use of AI is pretty interesting in here. If you look at -- we have capabilities in there like chat bots and IVR, which is interactive voice response. And people could have easily built those themselves on top of Lex. But they love the fact that we made it a push-button feature where they could just get IVR or they could just get chat bots. And they say, "Can you do more of that?" And so we looked at an area that we hear a lot of feedback from customers that they wish was easier. And that is doing analytics around their calls. And people say, "I want to store all these phone calls, I want to transcribe them to text, I want to actually be able to do search on them. I want to actually know what's in them without having to read every single one. I want to know if there's positive or negative sentiment. And I'd like an alert if there's some kind of problem."

And our answer to date has been, "Sure, that's really easy. What you do is you store all the data in S3, you use Transcribe to transcribe the audio to text, use Elasticsearch to make it searchable, you use Comprehend to do the natural language processing. And you use SNS or Simple Notification Service to create alerts." Customers say, "Well okay." Some customers say, "Great, that's what I'll do. And they were on their way." But we have other customers in that second segment who say, "Ugh, like, okay. But can you just make that easier?" And so I'm happy to announce today the launch of Contact Lens for Amazon Connect, which is machine learning-powered contact center analytics for Connect.

And so what Contact Lens does for you is you can activate it with a single click. And Contact Lens starts to transcribe all this data and analyze it automatically. And for each call, it'll provide a full text transcription, it will tell you the positive or negative sentiment nature of each contact. It will capture things like if there were long periods of silence, which often means an agent doesn't know the material or maybe there's some unhappiness or if people are talking over each other, which also, by the way, usually is a bad customer experience. Then it lets customers search against the transcriptions by keywords, by specific phrases, by sentiment, by things like long periods of silence or people talking over each other or by multiple dimensions, like give me all the contacts that had negative sentiment that talk about shipping delays. You can create dashboards so you can show the status of your overall contacts and your level of compliance against the SLAs that you set. And so it is a very different capability. In mid-2020, we'll also give you the ability to see these transcriptions happening in real times. And they'll call out for you whether or not you have a problem so that you can take action in the middle of that contact. Very excited to give customers this capability. Some of the customers that we have shared this with privately have said, well, this is awesome. I love that you can make sense of my customer service contacts. But if you can provide this level of understanding for all my customer service phone calls, why can't you provide this level of understanding with data that lives inside my own enterprise? And if you think about it, if you have an enterprise that's anything like ours. And Amazon is a reasonably technically savvy enterprise, there is all of this data that you have internally that lives everywhere. It lives in SharePoint files, it lives on Internet. It lives in file systems all over the place. And it's really hard to do the work to unite all these silos and then actually build some kind of index.

Then if you were lucky enough to unite them and build an index, then to build search that actually is useful is quite hard because almost all the search on this type of internal enterprise data is keyword-based which often doesn't answer the questions you want. Like if I want to know where is the IT help desk in the re:Invent building, I'm not sure exactly which keyword to say.

And so on top of that, because it's hard to unite an index and build sophisticated search, the results you get when you search your internal data in enterprises is gobbledygook. It's like a bunch of these long lists of links that have low relevance. And they don't really help you find the data. All that data inside your enterprise remains hidden and frustrating. And so this was another very obvious place that, frankly, our developers and customers pointed out to us that we should try to help solve. And that's why I'm excited to announce today the launch of Amazon Kendra, which is a new service that reinvents enterprise search with machine learning and natural language processing.

We're super excited about Kendra, we think it's going to totally change the value of the data that you get from all the data that lives inside your enterprises. And to give you an idea of how it all comes together and works, I'm going to welcome back to the stage, Dr. Matt Wood.

Matt Wood {BIO 18000850 <GO>}

Thanks again, Andy. So Amazon Kendra allows you to completely reimagine your internal enterprise search using machine learning but without requiring your teams to have any machine learning expertise. And I'm going to give you a quick example of how to set it up and some of the key capabilities.

So with Amazon Kendra, you can set it up entirely through the AWS console. You get started very, very simply by configuring the data sources inside your organization. These are the silos of data. And we have custom-built connectors where you just have to provide your credentials. And Kendra will go ahead and index and inspect all the data inside those silos.

Next, you simply provide optionally a set of FAQs, some frequently asked questions. These are really common in things like knowledge bases and support workloads or just new hire documentation when you have a set of questions and a set of answers. So you can provide those and the locations of your documents from S3. And we index those separately using our machine learning models under the hood.

The next step, you just sync and index your data. Kendra goes ahead, pulls in all of your data and builds an index. And we're not just indexing the keywords inside the document here. We're using machine learning, natural language understanding in order to be able to identify the concepts and the relationships between the documents based on the text inside those documents. So unlike the worldwide web, where you can rely on the structure of HTML. And you can rely on the structure of links between those documents, that structure doesn't exist inside enterprise data sets. Instead, you just have tons and tons of unstructured data set across all of these silos. And Kendra can pull all of that in, understand and relate the concepts inside that data and then build an index which you can query using natural language. And you can set Kendra to automatically refresh that index whenever you want.

Next, directly in the console, you can test and refine your queries. So you can do test searches and then refine the results in real time. So for example, if you have a query for sales reports where you want the most recent sales report to pop up to the top, you can just drag a little slider and away it goes.

Finally, you can deploy it. So Kendra comes with a pre-built web application which you can just drag and drop and host on your Internet or you can cut and paste code from Kendra, will automatically generate the code for you. And you can drop it into your existing internal applications. And it will just start to integrate and search with all the capabilities that you expect, including things like typeahead prediction.

So let's take a look of how this search performs. So here is an old school (junkie) search. Here, we're just doing keyword matching. This is the current kind of state-of-the-art. And you can see a question as simple as where is the IT support desk, returns just a ton of low-quality, spurious results, which don't actually answer my question. So now I have to go in and click on these links and try and figure it all out for myself, I haven't really saved any time by doing the search.

With Kendra, because we're looking inside and understanding the relationships and the content itself, we can answer natural language queries such as where is the IT support desk with a real answer. So we can say, "It's on the first floor" and point to the document where we got that from.

We can do other things as well because we understand the concepts inside the documents. So what time is the IT help desk open? Here, Kendra understands you're searching for time. It understands the concept and the relationship of times, dates and places. And so we can pull up from the document the answer to your question, "It's open from 12:30 to 5 p.m. daily."

Another benefit of using machine learning under the hood is that just by using the service, the machine learning models can get better and better and better. And so without your teams lifting a finger and going into all of the machine learning models, we can take feedback, smiley faces and upside down smiley faces to show whether those relationships and whether those results were useful. And we'll also automatically track what links your end users are clicking on and use that to continually improve the models under the hood without you having to do any of the custom labeling or training yourself.

So Amazon Kendra is incredibly easy to set up. It allows you to combine your data sources to provide accurate answers to natural language queries and, using our machine learning models, continuously improve with no machine learning expertise. It's incredibly exciting to be able to reinvent search with machine learning.

And with that, I'll hand it back to Andy. Thanks.

Andrew R. Jassy {BIO 15111610 <GO>}

Thank you, Dr. Wood, I am very excited to see what you all do with Kendra.

So when you look at machine learning end to end, as we said earlier, sometimes people trick themselves into thinking that machine learning is like a service, a single service. That's not what machine learning is. You need the right highly secure, highly reliable, fully featured data store with the right access control, the right security, the broadest set of analytics and really robust offerings at all 3 layers of the machine learning stack which we believe that most companies with modern technology capabilities in the future will clearly operate at all 3 layers of that stack. That's what machine learning is and what's needed. Nobody has that set of capabilities collectively like AWS, which is why twice as many companies are using AWS for machine learning as anybody else.

Now when you think about transformation, I mentioned earlier, there are things that you have to do to transform yourself, there are things you have to do to transform to meet new technical opportunities and situations. But one of the things that we notice when you get through these first 5 pieces of transformation that we've talked about is that, oftentimes, people will get addled if they can't figure out how to move every last workload. If they have some workloads that must remain on-premises somewhere, not in the cloud, they sometimes stall their plan in how to make this transformation. It's easy to run or to hide when you have angst about a big change that you've got to make. But it doesn't accomplish very much. And we see this type of activity and thought process sometimes in various companies who have the hard task of making this big transformation that if they can't figure out how to move every last workload, they don't know how to move forward sometimes. And so that's something that we have tried to help with over the last number of years and are continuing to make that easier for customers. So I'll mention 3 of them.

The first is really around customers who say, "I'm moving the overwhelming majority of my applications to the cloud. But there are some workloads that have to stay on-premises, maybe because they have to be close to something like a factory. How can you help me do that?" And this is kind of how we got started with the very unusual and deep partnership that we have with VMware and the joint offering that we spent a bunch of time together working on and launching called VMware Cloud on AWS, which allow customers to use the same software and tools they've used to manage their infrastructure via VMware on-premises. But also to use it to manage their infrastructure in AWS. And this has been something that customers have been very excited about. This is the only managed service that VMware runs of this sort like this and has a fair bit of traction. We have now 4x more customers than we had a year ago at this time, 9x the number of VMs that were there. And a number of companies that you know of that are using it, like Cerner. And Accenture and PennyMac and S&P Global and (Scripts) in the state of Louisiana. But customers have appropriately said, "Well that's great. I love that I can use the same tools I've managed my on-premises infrastructure to manage AWS. But that makes it easier to move applications to the cloud. What about those applications that I told you guys I have to keep on-premises for a while?"

And this was very much something that we were thinking about over the last few years because there was a model out here to try and solve this issue that we've heard from customers didn't work for them. And that's because they provided this solution that was different APIs, different tools, different control play and different hardware. And it was really hard for customers to use. And it's not that surprising. When you're taking 2 things, they're as different as on-premises and the cloud and then try and connect those 2 things together with a clunky bridge, you end up with something that's clunky and hard to use.

And so when we thought about this problem, we took a different approach, we thought about it less as building this clunky bridge between these 2 different things and more about distributing AWS on-premises. And that's why we announced last year in this keynote that we were building something called AWS Outposts. And so Outposts are racks of AWS servers that we deliver to your on-premises data center. And it's got AWS compute and storage and database and analytics. And you decide what composition you want and we'll deliver it and plug it in for you and set it up and maintain it and patch it. So it's not a lot of work for you. And we announced that it was coming a year ago. We've had a lot of customer conversations about things that were really important to you that we build inside of it. And I'm excited to announce the general availability today of AWS Outposts.

And so Outposts will come with EC2 and EBS and ECS and EKS and EMR and VPCs and RDS. And we'll be adding S3 in the first half of 2020 and a bunch of other things over time. But this makes it such that you can now run those workloads that need to live on-premises because they have to be close to something with the same AWS APIs, the same AWS control plane, the same AWS hardware and tools that allow you to leverage that learning and seamlessly connect with all your other AWS applications in our public region. So it's really easy to get started. You just go to the console, you set up an Outposts, you decide what composition you want of computer storage or database, we deliver the Outposts to your door. We'll install it. And we'll handle all the maintenance. Then once it's plugged into power or network, you'll be able to see your new AWS Outposts in your AWS management console and be able to use it to provision resources into your Outposts. So very excited that this is available today.

It comes in 2 variants. If you're somebody who is used to and wants to use the AWS APIs and control plane to operate your Outposts alongside what you're doing in the rest of the AWS public regions, then you'll use the AWS native Outposts. And that's available today. If you're somebody that wants to still use the VMware control plane, like you're using for VMware Cloud on AWS as a number of customers are, you'll use the variant of Outposts that's our VMware Cloud on AWS Outposts. And that will be available in the early part of 2020. So very excited to give this to you today.

Now we solve for this issue of I have certain workloads that need to live in my on-premises data center and can't move to the cloud. But there's a second barrier that we're hearing from a number of customers. And it's particularly larger organizations who say, "I have end users in a particular geography that have workloads that are latency-sensitive where they need single-digit millisecond latency and where I don't have a data center or I do have some kind of colo or clunky GPUs under my desk that I don't want to manage anymore. What can you do about that?" And that's a very interesting issue as well. If you're a cloud provider like us because it's very expensive to launch these mega regions like we have in 200 or 300 cities around the world. We have a lot more coming. But I don't know if we've really contemplated 200 or 300 more. And so we thought about this issue. And we thought about can we actually provide a solution? Take the typical examples, these are real examples of if you're a media company in L.A., where you do content creation or you do video games, those workloads as they're building them need single-digit millisecond latency. Or take a company in New York or Switzerland who are in financial services, they have to be close to market data, they need those single-digit millisecond latency. And so we thought about, is there a different type of construct we can provide you that solves this issue at scale? And I'm excited to announce a brand-new type of AWS infrastructure deployment called Local Zones, which is a new type of infrastructure deployment that places compute storage and database services close to large cities, starting with our L.A. local zone available today by invitation.

And so Local Zones have compute and storage and database available to you. And we kind of went back to our roots a little bit as we were solving this problem. We said, "If you think about the history of AWS, we have built in these highly flexible, low-level building blocks that you can build a lot on top of." Think about how many services, both AWS services and your services are built on top of things like S3 and EC2. And as we are worrying and thinking about this type of problem with Local Zones, we realize that Outposts was a really useful low-level flexible building block.

And so we've taken Outposts. And we've done some innovation and variance to it. And now for customers, we couldn't bring an Outposts to their on-premises data center in those areas because they didn't want on-premises data centers. We've built Local Zones in metro cities that are buildings that we manage that have Outposts in them with compute and storage and database and analytics so that you can have single-digit millisecond latency for your end users in those metro cities where they need that latency to get their job done. The first one is available in L.A. today. And you can expect more from us moving forward.

So we solve for workloads that have to stay in on-premises data centers, we solve for workloads in certain geographies where your end users need single-digit millisecond latency there. But you don't want to have data centers.

The third barrier that we're increasingly hearing from customers is that as they have more and more mobile and connected devices all over the world that have to be connected to the stealth network, how can they get that type of single-digit millisecond latency there as well. And with the promise of 5G, people have become very curious about this. They've started to believe this is possible. But like most new technology, this was true with cloud and with big data and with machine learning and with IoT, it's becoming true with quantum computing and it's true with 5G as well, there is a lot of hype and a lot of misunderstanding about what the technology is. So if we want to try and figure out collectively how we can leverage 5G to give those types of customer experiences, first, we have to really understand what 5G is and what it does and then how we can leverage it.

And I can't think of somebody better to explain to all of us how 5G works than our very close partner the CEO of Verizon, Hans Vestberg, please join us.

Hans E. Vestberg {BIO 1930652 <GO>}

Andrew, great to be here. Thank you.

Andrew R. Jassy {BIO 15111610 <GO>}

Welcome, Hans, appreciate you being here. Let me start with a question which is what the heck really is 5G? How is it different? And why is it so much better than 4G? What does it matter for all of these folks?

Hans E. Vestberg {BIO 1930652 <GO>}

It matters a lot for all of you in this room and it matters a lot to you. 5G is so different than any other G we have ever seen before. Think about 2G to 4G, was basically you took a 2G phone with SMS and voice. And you have today a 4G phone with great experience, if it's a Verizon phone, you can stream it. 2 different capabilities, speed and throughput. Those were the only 2. In 5G, it's actually 8 capabilities. We call them the 8 currencies. The 8 currencies, I'll give you some examples, just to understand how big difference is. So the speed is going from -- if you have a 4G phone today, 40, 60 megabits per second to up 10 gigabit per second for an individual device. And if you have a 5G phone in this room right now, you have 1.2 to 1.8 gigabit already because we have 5G in here. So that is speed. Throughput is going to be terabytes per square kilometer. Today, we do gigabytes per square kilometer.

Latency, which is super important for the services that you are developing, we're going from maybe 40 to 80 milliseconds down to 10 milliseconds on latency in the network. Then you can combine all of these. But you also have how many devices you can connect. Today, we can connect 100,000 devices per square kilometers, in 5G it's 1 million. So it's much more than people. So it's devices. And you can go on with other currencies. But I think the most important is that when you can slice this. And you give them to individuals, applications and things, you have a transformative technology that's going to transform consumer behavior. It's going to transform businesses. It's going to transform society.

You need to build this in a certain way. You need fiber to basically load the radio base stations, you need high-frequency spectrum. You need SDN, software-defined networks. You need to virtualize the network. And you need a lot of real estate at the edge. Verizon builds all of these 8 currencies, not everyone going to do it. We build it right, as I say, because we want to give this to all our customers. And already now, we have launched 5G home in 2018, which is a fiber-rich home broadband. We have launched 5G mobility. We did that in April. We have now 18 cities up, we want to have 30 by year-end. And of course, we are continuing with a lot of new innovation. But this is just the start of it. So the 8 currencies is what matters to everybody in here because that's what it can develop on. And that's the new things we're bringing with 5G.

Andrew R. Jassy {BIO 15111610 <GO>}

That's pretty awesome. So people won't have to be tethered anymore to the Wi-Fi or the lower-performing LTE networks?

Hans E. Vestberg {BIO 1930652 <GO>}

I think that, first of all, with our spectrum position with millimeter wave, with these big frequencies, you can get enormous throughput. You can get to 10 gig. And you can do the latencies of 10 milliseconds. So of course, that's why we can do it. But secondly, the connectivity and the speed is just 2 things. The other thing is we can with the 5G now bring the processing out to the edge, because we have a virtualized network, where we can bring it out. We call that the mobile edge compute for 5G. That's really what we can do. And here, the innovation of consumers, low latency, immersive experience, machine loads that are heavy, all that can happen at the edge right now because of the virtualized network and the 8 currencies.

Andrew R. Jassy {BIO 15111610 <GO>}

Yes. That makes sense. Well you could tell that 5G is pretty compelling. And if you think about -- as we thought about what that means for our AWS customers, if you want to have the types of applications that have that last-mile connectivity but that actually do something meaningful, those applications always need some amount of compute and some amount of storage. And what they've done in the past is they've reached out to AWS to go get that compute and storage. But the problem is there are so many hops along the way, like you have to go from the device to the cell tower to the city aggregation site to the regional aggregation site to the Internet and then to AWS and then back. And the types of applications that are most excited about using 5G, these are things like machine learning at the edge or autonomous, industrial equipment or smart cars or cities or augmented or virtual reality. They can't afford and don't want that round trip back and forth. And so what they really want is they want AWS to be embedded somehow at these 5G edge locations. But think about it, if you're a customer, what you want to do, you can't walk up to a telco and say, "Can I be at your edge?" Then, because no one telco can serve the entire world geographically, you'd have to go to lots of telcos and say, "Can I be at your edge?" Then you'd have to manage all the work between these different telcos and how they manage it with some kind of abstraction. And you'd still want all the same API tools and APIs and control planes and things of that sort. And so this has been a hard problem for people to try and solve.

About 18 months ago, Hans' team and my team started working on this and have collaborated very deeply. And excited to announce a new AWS service for you now called AWS Wavelength which allows you to build applications that deliver single-digit millisecond latency to mobile and connected devices with AWS compute and storage embedded at the edge and 5G.

And so what we've done is we've embedded AWS storage and compute, at the edge of 5G networks. We've started with Verizon who was our collaboration partner and really a true innovator in this space. And now for the sensitive portions of your application where you want the single-digit millisecond latency, you have way fewer hops to actually get to the compute and storage. You go from the device to the aggregation, the city aggregation site. Then right at that 5G city aggregation site is AWS. And so it's a much better experience. You don't have to worry. We're launching. Our first launch partner is Verizon. We'll also have KDDI and SK Telecom and Vodafone as part of this and more over time. You don't have to worry about figuring out how to manage across all the different telcos. We've built an abstraction that makes it easy for you. And as I said, we started with Verizon who's often the leader and the innovator in this space. And tends to invest most in this network.

Hans E. Vestberg {BIO 1930652 <GO>}

Thank you. Thank you. Now for us, this is a historic day. I mean, as I said, we were first with the 5G home. We were first with 5G mobility in the world. And of course, to be here today to talk about we are now bringing the 5G mobile edge compute out to the edge, we call it Verizon 5G Edge. Then with the collaboration of our engineers that has been working together for 18 months, we virtualize our networks, all the way from the radio to the packet core. And then bring in wavelength there in order to create platforms you'll -- all of you out there to actually start innovating on the 8 currencies, low latencies, massive throughput and all of that. And I think that it's just a massive moment for us at Verizon to actually do this and lead this sort of movement in the world with a lot of transformation with this new platform. We already are in preview in Chicago. So we have already customers up using these technologies. So it's much more than just the buzzwords. We actually have in preview. And we have NFL there. We have Bethesda, which is a gaming company that is using the low latency. But over time, we will be able to do slices with all these currencies, together with AWS Wavelength, where we can move the loads, of course, all the way up to the higher levels of the data centers down to the Wavelength. And that's a beauty because that's going to be a game changer. And I think that everyone here should be extremely excited for it. We at Verizon are extremely excited to collaborate with the best and greatest cloud company in the world and bring these to you as innovators out there.

And if you really want to see it today, you can go into our booth because Verizon has a booth in here, where you can actually start to experience these 8 currencies. But as the year progress in 2020, we're going to deploy much more real estate when it comes to edge compute together with AWS and see that we are bringing this to all of you. So you can innovate for your customers and partners. So it's a great day.

Andrew R. Jassy {BIO 15111610 <GO>}

It is. In case it's not obvious, we're pretty excited about giving this and we think it complete -- we think it pretty dramatically changes what customers are going to be able to get done. We couldn't have chosen a better collaboration partner. It's a deep partnership. We love working with Verizon and with your team, Hans. Thank you very much.

Hans E. Vestberg {BIO 1930652 <GO>}

Thank you.

Andrew R. Jassy {BIO 15111610 <GO>}

It was great. Thank you.

Hans E. Vestberg {BIO 1930652 <GO>}

Thank you, man.

Andrew R. Jassy {BIO 15111610 <GO>}

So we talked about, as you're making this big transformation, sometimes what happens is that companies get stalled when they can't figure out how to move every last workload. And that's what we spent a fair bit of time trying to solve. And we'll continue to work on trying to solve for you moving forward. But these are 3 big barriers I think we've been able to help you knock down. How do I deal with my applications that have to stay in my on-premises data center but want to work seamlessly with the rest of my AWS applications; how do I deal with end users and geographies where they need single-digit millisecond latency but I don't want data centers, again, that works seamlessly with the rest of my AWS applications; and then how can I leverage this incredible innovation in 5G to have single-digit millisecond latency with all my connected devices and the latency-sensitive parts of those applications, those are 3 big barriers that we have knocked down today for you.

So I'm going to close with just a few comments. I think that when there's a big change happening. And I think that the change that's happening with the move to the cloud is the most titanic shift that we've seen in technology in our lifetimes, it's sometimes hard to think about how to handle it. And I think that a lot of people will tell you that they love and embrace change. But I would say that my experience is that that's not necessarily true. I think a lot of people get nervous about change. And they don't know what it means for them and whether they have the skills to be successful in that change. And what it means for the scope of their job and all the things they spend a lot of time working on and the suppliers they built relationships with.

And as such, a lot of times, when there's a big change and transformation like this, the first reaction is to dismiss it. Then when it becomes hard to dismiss it because people are moving to it because there's value, then the reaction oftentimes is "Well we can do it better. We can do it less expensively. We can do it more performantly. Then when that doesn't appear to be true, a lot of times the reaction is just to try to slow roll it, do just enough to nod at it, that it looks like you're actually paying attention to it. But the problem is if you dip your toe in the water for long periods of time in transformations that radically change industries, you find yourself at the tail end of a big shift and suddenly way behind. And sometimes, it's pretty startling how far behind you become how quickly. I mean the history of business is littered with companies that did not adjust to big technical transformations and were left in the dust.

The reality is with what the cloud offers you, it gives you a once in a lifetime chance to totally reinvent the customer experience, to totally reinvent your business and to build things that were never possible before. And that to me is an opportunity that all of us have that is unlike any other in our lifetime and probably will be unlike any other moving forward. And so the opportunity is right here right now. Take it. We'll be here every step of the way to help you. And I hope you have a great rest of the week at re:Invent. Thank you.

This transcript may not be 100 percent accurate and may contain misspellings and other inaccuracies. This transcript is provided "as is", without express or implied warranties of any kind. Bloomberg retains all rights to this transcript and provides it solely for your personal, non-commercial use. Bloomberg, its suppliers and third-party agents shall have no liability for errors in this transcript or for lost profits, losses, or direct, indirect, incidental, consequential, special or punitive damages in connection with the furnishing, performance or use of such transcript. Neither the information nor any opinion expressed in this transcript constitutes a solicitation of the purchase or sale of securities or commodities. Any opinion expressed in the transcript does not necessarily reflect the views of Bloomberg LP.