10242024 Berlin Germany AI Tour Keynote Satya Nadella

CEO EMEA Tour
AI Tour Keynote
Satya Nadella, CEO & Chairman , Microsoft
Jared Spataro, CVP Modern Works & Biz Apps
Hall B, Level 2, City Cube Berlin, Messedamm 26, 14055 Berlin
Berlin, Germany
Thursday, October 24, 2024

SATYA NADELLA: Good morning. It's fantastic to be back in Berlin, back in Germany. I was just thinking about this, this morning, coming here. Microsoft's obviously been in Germany for now four-plus decades, and I've been coming back here for all of my professional career at Microsoft over 30 years across all the different tech platforms that I've had the privilege to participate in.

In the early days, it was all about the PC, client-server. In fact, I spent a lot of time in Waldorf doing the R3 port to SQL Server. And then, of course, there was the web-internet era, and then mobile-cloud, and now here we are in what's the fourth big platform shift that at least I've witnessed, with AI.

It's always exciting. Those beginning years of anything new I think is just an exciting time for all of us in tech, and, quite frankly, for all of the folks who are looking at this technology and its impact in the real world.

It's always helpful to start with a deep understanding of the underlying force. In fact, when I go back in '92 when I joined the company, it was very, very clear that there was a force called the Moore's Law that was driving the entire computing industry. In fact, I remember going to PDC, the Professional Developers Conference, in '91 and really being able to see what Moore's Law was going to do to server computing. In fact, NT was not even there at that time, but it was clear as day that the Intel architecture was going to basically change even what was the backend computing. It was already taking shape on the PC, but client-server was going to just happen, and it just literally happened in five years, just because Moore's Law was doubling every 18 months.

Similarly, when it comes to AI, we have what we describe as scaling laws. Just like Moore's law, it's an empirical observation, quite frankly. You see that you now have performance doubling every six months.

The new equation that I like to talk about as the new currency for any country, any society, I think going forward, will be tokens or performance per dollar per watt. If you think about it, that's what you are going to really say is the real measure of how are you progressing. We are now seeing that doubling every six months.

It's a combination of things. It's compute, scale, but it's also really how you use data, both for pretraining and post-training, what changes you make to the algorithms and the

algorithmic architectures that you use. That combination is really explosive; you have these exponentials.

Now, you could say this started, in fact, way back in 2010 with deep learning, but in 2018, 2019 is when it started inflecting, especially with transformers of natural language, and now you have the manifestation of this scaling law in three dimensions.

The first is essentially for the last 70 years, one of the dreams of computing was, can we have computers that understand us versus us having to understand computers? The natural user interface has been something that we've always been pushing for, and finally, I think we have it. With these new AI models you have speech in, text in, video in, image in and out. The fact that you have a multimodal input and output is going to transform pretty much every computing interface.

Second thing, in fact, the other 70-year dream was every day in digital technology, you digitize more artifacts about people, places and things, and then you reason about them. In fact, when I go back to what was database technology, it was all about using relational algebra to reason about people, places and things. Now we have new algebra. It's the neural algebra to be able to make sense and find patterns, do predictions on people, places and things.

Then the third thing is you can also give it more context, more memory.

If you put these three things together, you can build this very rich agentic world, or AI world, where you will have personal agents. You will have agents that are working on behalf of your team, or agents that are working on behalf of your business process or function, marketing, sales, finance, across organizational agents. That's the rich tapestry of agentic playing field that's getting built on top of all the digitization and systems that have come before.

Now, of course, this is all wonderful; it's all technology. The real question, of course, for us, that grounds us is, how do we translate all this into real-world impact? Our mission is to be able to empower every person and every organization, starting here in Germany and around the world, to be able to use this technology and to really make a real difference in a small business's productivity, or the Mittelstadt's productivity here, or how do you really improve the efficiency of a public sector organization or a large multinational company out of Germany becoming more globally competitive, education outcomes, health outcomes. That, at the end of the day, is what all this technology is about.

Now what I want to do is talk about three platforms because Microsoft has always been a platform company and a partner company. We build platforms so that others can build more technology. I want to talk about the three platforms we are building: Copilot, Copilot Stack, and AI platform and Copilot devices.

Starting with Copilot, the best way to conceptualize Copilot is think of it as the UI for AI. That's the organizing layer, in fact, for this agentic world, which is going to be very, very

important, because ultimately, even in a world where you have millions of agents, you need a UI to deal with that agentic world, to make sense of that agentic world. That's really what this Copilot interface is all about.

Now, one of the other things is the Copilot interface, of course, is already helping you, especially in the work context, retrieve information from the web, but also retrieve information about your work. All of your work artifacts are something that you can really reason about.

But it's not just the fact that you're able to recall information or data; you're able to create new AI-first artifacts. In fact, the first one that we created was this Pages. Pages is pretty cool. In fact, the metaphor I use sometimes is, I think with AI and I collaborate with my colleagues. That's a simple way to think about how the new work and workflow gets shaped.

Pages is this new AI-first artifact in which I can take some of the information I've retrieved from the web, from my work databases, put it in, have even AI operate it and manipulate it, and then share it with my colleagues, who can then, of course, in real time, even collaborate with me on it. That is just a work system going forward.

Now, we don't stop there. In fact, you go to the next step, you can take a Copilot product or the interface and extend it, using Copilot Studio to build agents. The extensions of Copilot are agents.

Copilot studio is this low code/no code tool. Just like back in the day, we would take Excel and create spreadsheets, add formulas and macros, similarly creation of agents is similar to just that same process, which is you use Copilot Studio to create agents.

In fact, the data that is available to you is this rich fabric. You have everything inside of Microsoft Graph. In fact, the most important database in any organization is the database of the people, their relationship with other people inside the organization, all of the work artifacts, whether they're documents, project files, all of that, emails, messages and teams. Everything is now available to you in a graph structure to be able to interrogate and make sense.

Of course, you have all these connectors to all these other business applications. That's what Dataverse is. Whether it's your SAP backend, your CRM backend, or what have you, you can get all that data as well, and then, in fact, all the data you may have underneath your own line of business applications, or the analytics work you have done in Fabric. You now have the ability to build these agents on top of this data and dock them with Copilot, which is the UI for AI. That's, I think, a way to think about it.

In fact, the ambient video here is a simple example. Let's say I want to build a field service agent. All I've got to do is go to Copilot Studio, give it a little system prompt, and then after that system prompt, I anchor it. I anchor it with some data and knowledge. One piece of knowledge here is a SharePoint site, in which there's a bunch of documents

around field service. And then a system of record here happens to be a Dynamics database, and so I can say field service in Dynamics.

That's it. It's like, literally point and click, where I've been able to give it a system prompt, some data sources, and I have a field service agent now that is available to me inside of Copilot.

That type of experiences are what all of us will create, and that, I think, is the crux of it. Sometimes we mystify AI and agentic creation. It's like creating a spreadsheet. It's no different than that. Nobody goes and asks how many spreadsheets exist inside my organization. Maybe millions. Similarly, the way to think about the new work, workflow and work artifacts are we're going to create agents just like how we created back in the day documents or spreadsheets. It's that simple.

The impact of this is already being felt. Right here at Microsoft, for us, we're seeing double-digit changes in terms of productivity, in fact, growth. When I think about sales and marketing, we are seeing 12%-plus, 10- plus growth numbers because of these Copilot, Copilot agents.

We're also seeing all the self-service applications. When you think about whether it's in HR, whether it's in IT or customer service, we are seeing fantastic efficiencies in throughput, cost savings in legal and finance.

These are tangible. These are things that would have showed up in compounding at probably 2% to 3% every year. They're showing up as double-digit numbers. This is not linear, even in terms of the benefits. The scaling law is on the training side, but on the inference side, you can think of this as showing up as double-digit gains in a year. This is a fantastic to se. It's not like we are waiting for this, or this is a POC somewhere and the results will come years in the future, but it's happening as we speak.

That's also being manifested. This is not just Microsoft data. At this point, there is enough evidence across the longitudinal studies that we have been doing across the usage of Copilot.

It is fantastic for me to even see all the examples in Germany, Bayer and BearingPoint and Eon and Porsche and SAP and Siemens and Juniper. In fact, one cool example I saw today was from IU, which is obviously International University, that's doing some phenomenal work in educating people. They built their assistant, Syntea, and they built, in fact, an agent, or a personal tutor, if you will, for the real world, for anyone who is going into university there.

But they have done a version of Syntea for work. In fact, think of it as an agent that docks with Copilot. They call it Copilot, Syntea Copilot, and it's really very, very well done, because it really then goes into my graph, looks at my role, looks at all the things I'm doing, the meetings I'm having, the work artifacts I'm creating, and creates, essentially, a

personalized curriculum for me to be able to do just-in-time training for the job that I have.

That's just an example of even what was in the past, let's call it corporate learning apps and corporate learning process being streamlined such that it is now being delivered just when you need it in an interface like Copilot. That's fantastic to see the progress there.

Of course, we're not stopping here. This next phase is going to be also about autonomous agents. In fact, we had, earlier this week, made two announcements. One is that same Copilot Studio, where you used it to build agents like that field service agent, you can now use the same design time to build autonomous agents.

And then, in fact, we also announced that we have 10-plus autonomous agents that we're going to launch, everything from sales qualification, supplier communications, to financial reconciliation. Think of all of these as business process tasks.

The thing to recognize is autonomous agents obviously work on their own, but from time to time, they need to call on us to help them. That's called exception handling. When they need to have an exception handled, the way they do it is by raising the exception back to a Copilot workflow.

That's where I think this is not about autonomous agents on one side and Copilot on the other side. They're all connected. You need a Copilot as the UI for AI. You will have agents that come to Copilot, and you will have even autonomous agents that are working, but from time to time, will raise exceptions and dock with Copilot. That's the entirety of the ecosystem.

To show you a little bit of this in action, I wanted to invite up on stage my colleague, Jared Spataro. Jared, come on up.

JARED SPATARO: Thank you, Satya.

Good morning, everyone. Great to be here with you. I'm sure many of you are familiar with McKinsey and Company, one of the world's most successful management consulting firms.

At McKinsey, client experience literally is everything, and they're always trying to improve. You can imagine our delight when they decided to work together with us, using Copilot Studio to create an autonomous agent to improve what they call the client onboarding process.

It all starts with an incoming email from a prospective client, like the one you see behind me. Now, traditionally, these types of emails had to be received and parsed by humans, then going through the contents and trying to figure out exactly how to route this request for a new project.

But no longer. Now they have created an autonomous agent that springs into action as soon as an email arrives. It goes through a series of steps, essentially parsing the email, looking for the most important details and pulling them out. Then it matches that to an internal taxonomy, and even goes into a skills database to find the right McKinsey partner to take the next step. When it has everything that it needs, it takes that information, and it drafts an email to that McKinsey partner, summarizing exactly what it knows and exactly what it thinks should happen next.

Now, all that happened quite quickly. So let us just pause here for a moment to take a look at what it did. Essentially, we gave this autonomous agent some operational guidelines and then allowed it to take all of the information that came from the person through the email, parsing through the ambiguities of human language and matching this request up with the right McKinsey partner. It's pretty remarkable.

But that's only half the magic. The other half comes as we see how easy it is to create these autonomous agents. For that, we will move over to Copilot Studio.

Here in Copilot Studio, we see that we program up these agents, but not with the sophisticated programming language. We actually give the agent instructions, just like we might a human that was assigned the task.

Now, what makes the agent autonomous is what we call a trigger. In this particular case, the trigger is to listen for those incoming emails, as you see on the screen. But we can trigger off of many different events across many different systems, allowing us to have this agent work for us 24/7, listening closely to any set of signals across the IT estate.

We can also add additional knowledge or context. Here you see, for instance, there's a Word document, a SharePoint site and a database. But in the same type of way, you can add additional sources of information. Here we see examples from SAP and from ServiceNow, including other line of business systems that you might have.

Now, the final thing that we do with this agent is that we give the agent a set of actions. You saw those happening as it was parsing through the email. You can use over 1,500 connectors to take actions across all of your systems. It makes it incredibly powerful.

What is so differentiating about these agents, though, is that they don't scale like people. You saw an example of one email and the agent working through it, but in reality, agents like this can handle dozens, hundreds of different inquiries that come in all at the same time. Let's take a look here at the activity tab. Up top, you can see that, in fact, it has already completed almost 1,300 different sessions, with 33 in progress right now.

Now, Satya mentioned that every once in a while, these agents, well, they need a little bit of help; they get stuck. Let me show you what that looks like. We'll dive into the second one from the top here, and we'll see that it took exactly the same steps that we saw, but it does figure out at the bottom here, Retrieve Partner Information, that it's not sure what to do next.

Digging into that, we see that it has identified correctly the right partner, but then it realizes that partner is no longer at the firm. It can't send the email. Luckily, instructions have been given, and it knows that it should escalate up to a human manager to receive more instructions to know how to proceed.

For that, we are going to now go to that UI for AI, the Copilot, and we'll see what that looks like. Here at the bottom right, you will see almost right behind me a notification that pops up. When the user, the manager in this case, hits that, it gives them all the information they need to enter in the right person who should take this inquiry. Then with that information, the agent can complete its task, write the email, send it and log what it's done on the right-hand side.

Now, McKinsey has already put this through trials, and they have found that they can reduce lead time for what's happening here on onboarding a new client by 90%. They can also reduce administrative overhead by 30%.

But that is just the beginning. You see, we envision an orchestration layer of various agents, like this constellation of agents, that will help individuals, teams and entire functions literally rewire the way they get their work done.

Copilot Studio is where you create these agents. Copilot is where you go to interact with them.

Now, as you can imagine, we're incredibly excited about what you will do with this technology, and we've actually been working with some early adopters of the tech. Why don't we roll this video so that you can take a look at what they are already doing?

(Video segment.)

SATYA NADELLA: Thank you so much, Jared.

That gives you a real feel for this new platform for AI-driven business transformation. From Copilot to Copilot Studio to agents to autonomous agents at the end of the day is a platform that you use to be able to change, whether it's marketing or sales or customer service or engineering or any function in any role, and really get that next level of productivity.

Now, we'll move to the next platform, which is the Copilot Stack and the AI platform. The approach we have taken is everything that you just saw was built on a bunch of technology layers. What we are doing is we are exposing every layer of that tech stack as first-class developer services, so that you can build your own AI, your own AI applications.

It starts, in fact, with the infrastructure layer. When we say Azure is the world's computer, it starts by laying out this rich distributed computing fabric all around the

world. We have 60-plus regions, 300-plus data centers around the world. Right here in Germany, we have two regions which we are continuing to expand. In fact, we recently announced that we're going to invest 3.2 billion euros to expand our German footprint. It's infrastructure that has both traditional compute, as well as AI compute.

In fact, when it comes to AI compute, you see us again even there take a full stack approach. Innovation at the silicon layer, you have the best AI accelerators from Nvidia, AMD, our own with Maia, all coming to the fleet.

Then, in fact, it is fascinating to see, for example, some of the innovation we did with Maia with liquid cooling is what we will use, even with GB200 as they become available in the fleet. We're doing a lot on the silicon, system software, scale-up network, scale-out network. All of that is happening. We're optimizing it all for AI workloads, whether it's for training. Training is a very different class of workload because it's a synchronous data parallel workload. Test time compute or inference time compute is becoming more and more important.

Therefore we are doing a lot of system software to optimize for these new workloads. Just like, hey, back in the day, we would have done for databases or for any web servers or what have you, we're doing it now for AI jobs and AI workloads.

Now, the next thing after infrastructure is data because at the end of the day, AI is a compute meter that attracts data to it. In fact, there's gravity where data has to have nexus with your compute infrastructure.

What we're doing there is starting by making sure that you have the ability to take all your data – it may be in Oracle, it may be in Snowflake, it may be in Databricks – bring it to the cloud, bring it to Azure.

Then on top of that, we are building what I describe as AI-first data estate or OLTP. It's Cosmos DB. In fact, if you look underneath ChatGPT, it has all its sessions data. By the way, ChatGPT is a stateful application. Any AI application is not a stateless call to an API; they're very stateful application. Cosmos DB is heavily used. You have Postgres; you have SQL. All of these things are being built again for AI. Then we have the analytics infrastructure with Fabric.

Then, of course, one of the other important considerations when it comes to AI is data governance and compliance. Therefore, Purview is being built for all of the AI data workflows. The combination of OTP analytics plus Purview are all, again, being optimized for AI workloads, just like in the past, we would have taken data and optimized it for something like a line of business application being built on a database.

Now, the next thing, of course, is now I have my infrastructure, I have my data, and I've got to start coding. When you start writing your application, the first thing you want is app services. In fact, the fact is, again, go back to look at ChatGPT. The first thing you do when you start building ChatGPT is provision a lot of cores with AKS. You need a

Kubernetes service, and with Azure Kubernetes, that's sort of the first service. Then you may use Azure Functions. You may use lots of these traditional cloud native app services. That's the first thing you do.

Now, interestingly enough, we now are birthing a completely new app service for AI. In fact, Azure AI Studio, you should think of it as an app server for the AI age.

What does it start with? It starts with having the richest selection of models. Of course, it starts with all the OpenAI models. We're excited about even the new reasoning models with o1 and GPT-4o, still maintain their leadership position when it comes to frontier models. But it's also a full selection of models. There's OpenAI models, there's Llama, there's Mistral, there's Cohere's, there's open source, closed source. You have a very rich model catalog that is available to you as developers.

Then once you have your models, what you do is you start doing things like prompt flows or prompt orchestration. That's one. You do RLHF, reinforcement learning with human feedback. You do SFT or supervised fine tuning. These are all things that you will want to do on top of those models.

Of course, the way you do that is you also need to be able to take data and make it available to these models. For example, one of the most important application patterns is retrieval augmented generation, and that means you need to take your data and do embeddings on the data and then to be able to do retrieval on top of it. That's where Azure Search, which, by the way, again, is a service that ChatGPT itself for its RAG uses in a massive way. That's built in right into the app services.

You have the models; you do all the tuning. On top of that, you even have things like the helper functions for RAG. And then, of course, you're now ready to deploy. But before you deploy, you want to make sure that you have the guardrails and the safety.

That's where, again, we've built all the guardrails that work, by the way, across all model families, across all model versions. Because model versions will come fast and furious, but your application across the board has to be fairly stable.

That's what the app server does. You have traditional app services, you have the AI app server, and then, of course, you're finally ready to code away. That's where Microsoft from 1975 to 2024, there's one thing that we love to do, which is build tools for developers.

VS Code, GitHub, GitHub Copilot are the standard issues for every software developer out there in the world, and we are very, very proud of the progress we continue to make. In fact, I think next week is GitHub Universe, or the week after is GitHub Universe, where it's going to be fantastic to see all the advances at that conference.

In fact, the latest version of GitHub Copilot has this o1 preview in there, and if you see the ambient video, it's pretty cool. What it's doing is now you can use reasoning capabilities or these new models to actually optimize the algorithms in your code.

This one is a tokenizer which is used by GitHub Copilot itself. This is the tokenizer you use to do those byte pairs. What you have is AI being used to optimize the algorithms of the tools that are creating AI. In some sense, that's sort of a good way to think about how all of this is advancing.

In Germany, we have fantastic growth of the GitHub community. It's the second largest in Europe. We have 3.3 million developers in GitHub. It's continuing to grow 24% year over year. It's great to see that.

We have lots of customers already building first-class AI applications. This morning, I had a chance to meet with many of them. I saw developers from Lufthansa proudly talk about all that they're doing to transform their own operation, starting from the customer interface, back into their back office operations, using these AI agents and copilots that they're building on the Copilot Stack.

I had a chance to meet with developers from Otto. In fact, a developer there said something which was very cool, which is they said they want to turn every developer into a customer obsessed and customer focused developer, empowering them to be able to do more essentially. Because one of the things that AI does is it raises the floor and the ceiling for every one of us, so we can aspire to do more. The best way to think about aspiring to do more is to be more focused on what customer value you can create.

They're also into health care with Medgate. Just like Otto is using GitHub Copilot to make their developers much more customer focused, Medgate is also building Copilot for all of the doctor-patient interactions and transforming the medical systems.

Ottobock is doing some phenomenal work in prosthetics. They were describing to me how they're effectively creating new models to be able to get all of the data that allows them to do personalized prosthetics.

If you think about one of the fundamental challenges of any amputee is to be able to get a prosthetic that fits them, that's a big challenge today. In fact, for children, I know that personally, where it's very, very hard to get prosthetics that grow as children grow. To be able to have personalized systems where you're able to use AI to help, that's going to be a fantastic benefit to a very under-served part of our society.

I had a chance to meet with Siemens, and they are building out these industrial copilots, both for engineers as well as for operations. In fact, they're working with ThyssenKrupp to be able to deploy them and have a fantastic feedback loop, learning from real industrial engineers who are doing now this PLC programming, using effectively like what we did with GitHub Copilot for software developers, is happening for industrial engineers.

It's not just for industrial engineers, but even for the operations folks to be able to have a Copilot that helps them. It's kind of like the observability, as I learned today, for industrial companies. It's just unbelievable to see all of this.

Let's just roll the video of Siemens and ThyssenKrupp to just give you a flavor for the work.

(Video segment.)

SATYA NADELLA: It's fantastic to see this rate of diffusion. Again, going back when I came here in the early days of the cloud, cloud adoption was happening, but not at this space, where kind of like what's happening in GitHub Copilot, or what's happening in the Siemens industrial copilot is essentially similar technology being deployed at the same time across various domains, and that's fantastic, I think, in terms of the rate of diffusion.

Now, the last platform I want to talk about is the Copilot+ devices, or Copilot PCs.

Now, just like we talked about AI scaling laws transforming everything that's happening in the cloud, from the silicon up, the same thing is going to happen at the edge with devices.

In fact, it starts with silicon innovation. We're working with, again, Qualcomm, AMD Intel, to bring these new chips sets, basically which have CPUs, GPUs and NPUs, as much as 45 TOPS, so that at the edge, you have fantastic AI power.

Then on top of it, distributed computing will always remain distributed. It's not just the cloud AI fabric, but it's the cloud plus edge device AI fabric. You're able to write programs. Whether it's Copilot, whether it's Adobe, whether it's SAP, or whether it is your next video editor, what have you, you will be able to take advantage of the entirety of the silicon that's available across the cloud and the edge, and that's kind of the future of what every software developer can do going forward.

To just give you a little bit of a flavor for this next generation of Copilot+ PCs, let's roll the video.

(Video segment.)

SATYA NADELLA: Now with these three platforms, perhaps what's the most important consideration for all of us is trust in these platforms, because we know at the end of the day, if this technology has to be ubiquitous, used across all of our society, we need to have real trust in all this technology. Trust is not something you just claim; it's something that you have to earn.

The approach we're taking is pretty straight forward, which is, one, we have a set of commitments that we are making in terms of what we are going to do around security with Secure Future initiatives, the principles behind the privacy, principles behind AI safety.

Now, having made those commitments, the real key thing is to make progress in terms of capabilities we build across those commitments, and that's what you see us do. In fact, every time when we're talking about our product and features in products, we're talking about the progress we're making on the trust layer.

Even take security. I spoke a lot about all these new models. One of the first things you want to really do is when you're launching a model is to, in fact, think about its security and its robustness to attacks. The way to do that is to simulate, in fact, adversarial attack. We built an adversarial attack service essentially into that AI app server as an example.

Now, take privacy. One of the things is I already talked about Purview and data governance, but one of the issues with privacy is even when you are in the cloud, you want to ensure that your data, even during use, is secure. The way to do that is confidential computing. Now confidential computing is not just in the CPU, but it's coming to GPU. In fact, even Nvidia now supports confidential computing in the Azure fleet. That's just another example of making real progress when it comes to privacy.

When it comes to AI safety, I talked about the guardrails. We say, okay, LLMs have hallucinations. One way to think about groundedness of any model, any application is to, in fact, use AI to eval the groundedness of your model or your application. That's another service which we just launched in our AI App Server in Azure AI Studio.

These are things that we're doing specifically to make progress on the commitments we make around trust, and I think this is going to be the most important consideration for all of us, beyond just the features in any product.

Now, I want to close out by really talking about, ultimately, as I said in the very beginning, our mission is to be able to translate all this progress into real world impact. That means empowering every person and every organization right here in Germany so that they can achieve more. That's, I think, at the end of the day, what gives us the deep meaning in our mission.

Now, one way to really go about accelerating this is the human capital of this country, and helping the human capital get AI skilled. In fact, AI talent in Germany has been growing. This is LinkedIn data. It shows 71% growth. It's fantastic. Today, I'm really thrilled to announce the AI Skilling Alliance, where we are working with all the leading companies of Germany, as well as worker associations, to advance the workforce in Germany, so that everyone here in Germany is ready for this next generation of AI and the transformation that will be brought about across the length and breadth of German society and industry because of AI.

Thank you all very, very much, and hope to see you next time.

(Applause.)

END