09232024 Mexico City AI Tour Keynote Satya Nadella

CEO LATAM Tour
AI Tour Keynote
Satya Nadella - CEO, Microsoft
Rafa Sanchez - GM, Microsoft Mexico
Callie August (Demoer) - Sr. Marketing Communications Manager, Microsoft
Centro Citi Banamex - Hall D
Mexico City, Mexico
Tuesday, September 23, 2024

RAFA SANCHEZ: Hi. Good morning everyone. Welcome to our AI Tour Mexico. Thank you very much for being here. We have prepared an amazing agenda to show you the latest and the greatest of technology, especially on AI, security, cloud, data, applications and development.

Before we go there, please let me say thanks to all the customers that are here today with us, sharing with you the best practices, and for all the partners that are also here today with us, showing to you the use cases that we are running right now here in Mexico. I'm pretty sure that we are all here to hear directly from a person, from a leader that is shaping the business and the technology around the world.

Please join me in giving a warm Mexican Welcome to the chairman of Microsoft and CEO of Microsoft, Satya Nadella. Satya, the floor is yours.

SATYA NADELLA: Good morning. It's fantastic to be back here in Mexico, and at an exciting time for our industry. The last time I was here we were in the middle innings of the cloud transformation, and of course, it's now all about AI. This builds on everything that came before with the cloud, but it's always exciting in the tech industry when you're at the beginning of the next platform shift.

And so I wanted to share my perspective, what this means, and what it means to all the people who are innovating on top of these new technologies right here in Mexico. When you talk about AI, I think it's always good to ground ourselves in the fundamental driver of AI, which is the scaling laws. These scaling laws are like the Moore's Law, an empirical law that we observe.

When you think about what happened with Moore's Law, every 18 months we had a doubling of performance for the same price. We are effectively seeing something like that with these scaling laws.

In fact, it started in 2010 with the deep learning revolution, and it inflected again, perhaps with these transformers in the 2018 and 2019. When you observe what's happening with scaling laws, you are effectively seeing a doubling of performance, perhaps every six months.

If Moore's Law was fantastic, it's possible that these scaling laws, as they persist, are going to give us more abundance of compute and intelligence. That's foundational thesis behind this, or the foundational observation that drives everything else in AI. You can project forward five or ten years from now from now as to what will happen?

In fact, if you look at the prices of something like GPT models, in the last 18 months, they've dropped, what – 200 times? It's extraordinary to see this.

Now, all this is ultimately leading to more capability. It comes in the form of a complete new way for us to interact with computers. This multimodal input and output, I can talk to it. I can give it either vision or video, images or video. I can obviously give it text, and I can even have text output, video output.

And so this multimodal user interface, which is multi-turn and long in duration, I think it's going to fundamentally change how we interact with computers.

Second, you now have increasing levels of reasoning. All through digital history, we have reasoned about digital artifacts of people, places and things, with things like database technology and SQL queries. These were things we used in order to make sense of the digitized world, but we now have a new neural reasoning engine, and so these two capabilities, coupled with things like memory and context, are allowing us to build a real agentic going forward.

This agentic world has personal agents and has agents that work on behalf of teams and organizations, and so you can have your SharePoint agent, for example, you can have an agent for your sales team. You can have an agent for your marketing team. You can have even business process agents.

Of course, there are also cross-organizational agents. This agentic world is what is getting built on top of all of the digital substrate we are built to date. You can take the digital systems you can overlay on top of it, with this very rich tapestry of AI agents, starting with the personal agent.

Of course, it's not technology for technology's sake. Ultimately, we at Microsoft are grounded in our mission to translate this very powerful technology shift into empowering people and organizations everywhere to achieve more. That's really, at the end of the day, what I think everyone across this room also cares about. It's not technology, but technology's sake, what is its real world impact? One person at a time, one organization, one institution at a time.

That's our mission, and to that end, we are building three major platforms for this agentic world. The first is more on the personal agent side, which is Copilot Plus Agent. They will be both for consumers and for enterprises. We're also building out the Copilot stack, the Copilot plus agent stack. This allows every software developer, every organization, to be able to build their own AI agents for all of those organizations and business processes.

Lastly, we are also building Copilot plus agent devices. The first example of that is what we have done with Copilot+ PCs. These are the three platforms that we are bringing to market and bringing it to market in such a way that there is a rich ecosystem opportunity right here in Mexico as well.

I want to talk about each one of them, and so let's start with Copilot Plus Agent. The Copilot Plus Agent is all about bringing together a design system that spans everything from chat to Copilot Pages to all of the Copilots in all of the Office applications, whether it's Word, Excel, `PowerPoint or Outlook, and Copilot Studio, which allows you to extend this Copilot ecosystem.

That's really what we are building, and this is grounded, because Copilot is not just one thing, but you can actually ground and steer these Copilots for specific roles and functions. You can have Copilot for Marketing, you can have Copilot for Sales, you can have Copilot for Supply Chain. This means you are able to ground it in the data and the workflow for a specific role and a specific function.

Now, Copilot Studio is what allows you to be able to do all of that by really helping you build out these agent-to-Copilot interactions, and we are already seeing very rich productivity gains with the ability for anyone in any role to be able to get more done. The work, the work artifact in the workflow, is changing, and it's really getting to a place where the knowledge turns within an organization is becoming faster, just like how we have in a traditional industrial company, supply chain turns or inventory turns.

The question is, with knowledge work and cognitive work, can you have faster knowledge turns? That's what's happening. Customer service is more productive, marketing is more productive, finance is more productive.

I was able to see many customers, right here in Mexico, exploit this to drive productivity and real and creation of operating leverage. I had a chance to meet with folks from Cemex. They're deploying this technology broadly. They started with their customer interface. They're now moving it to operational processes that are critical to their plants. They're going to supply chain. They're also deploying for horizontal knowledge work throughout their enterprise Copilots. It's unbelievable end-to-end digital transformation, now driven through AI.

I had a chance to see the frontline agent that was being built by Grupo Bimbo, and what they're doing across all of their estate is kind of a sales agent. They have a complex set of products that frontline people now are able to interface with.

I saw a fantastic example of Capgemini using some of this technology to modernize Copilots in more of a GitHub Copilot example.

There are lots and lots of local examples of customers who are already driving a tremendous amount of productivity and transformation using these technologies.

Last week, we were very excited to announce the next phase of this Copilot Plus Agent journey. We talked about it as the phase two of the Copilot Plus Agent journey with something called Work, Web and Pages.

I think that this is a very transformative part of the design system. When you think about chat and being able to use chat as an interface to get knowledge from both the web and work that is fantastic.

Now, beyond that, I want to be able to share this with others inside my organization, and that's where Pages comes in. Each time there has been a change in the platform, there has been a new artifact that has gotten created.

We obviously created Office applications, as we know of it, when the PC first came out, and this was all transformed into collaborative canvases with the web, but we now have, for the very first time, an AI-first artifact in Pages, where I can think with AI and I can collaborate with my colleagues using Pages.

It's the idea of how work gets created or how work gets done and how the workflow changes, which is going to be unbelievably transformative.

To show you this at work, I wanted to invite up on stage one of my colleagues, Callie. Callie, come on up.

CALLIE AUGUST: Thank you, Satya.

I've been using Microsoft 365 Copilot in my daily work for over a year now, and it's remarkable how it's transformed by work habits. I personally us it for things like summarizing information, getting a jumpstart on a project, or finding that one rogue file, even how I prepare for a customer meeting here in Mexico City,

Copilot can get me up to speed in seconds. I simply have to prompt Copilot to find the customer prep call that I had before this meeting and ask it to give me an agenda and action items for my meeting this Friday.

Within moments, Copilot actually able to go find that meeting, put together a summary for me and build an agenda with clear action items, but this is really just a fraction of what Copilot can do.

Today, I want to show you a more collaborative, powerful experience with Copilot Pages. Let's see it in action.

Let's say you work at an EV charging company and there's a business opportunity in a new city you have never been to. I'm going to use LAX or this example.

To start, you want some basic information about LAX for your project.

Now, because Copilot is grounded in web data, it can reason over the entire internet and quickly pull in the relevant information.

In just seconds, you have a great start with answers to your question.

Now let's turn this simple response into a page.

Here, can keep working on your own, with your colleagues or with Copilot. You can even share a page just like you'd share a link to a Word doc.

Now, this research section is looking a little bit light, so let's tag in some of your colleagues to help fill this out further. Once you do, everyone can immediately start to contribute.

Now, while your colleagues pull in the latest industry news, you continue to do a bit more research to build out the business case. What you just saw was Copilot pulling in information from the web, putting it on a canvas and enabling you to collaborate with your coworkers, but what happens when you bring in your work data? I'm talking about your chats, emails, documents, meetings and more. That's where Pages really comes to life.

The first thing we're going to do is pivot from web to work. You need to get started on a plan for your proposal, but you want to be efficient and use some existing work. You ask Copilot to reference a previous plan and give you a head start on this one.

Copilot pulls in the deliverables, the work, back schedule and owners and formats at all as a table in your page. You are now ready to collaborate with your colleagues on the plan.

Here, you can see coworkers already updating the work back schedule and assigning owners.

Now, let's get to work on that proposal outline.

As a first step, you need to understand what the customer is looking for. Your coworker had an introductory meeting with the customer last week. Right here in Pages they have asked Copilot to reference the recording of that meeting and insert a bulleted list of requirements.

In seconds, Copilot was able to synthesize the discussion and give your team the key takeaways.

Now that you understand what it's going to take to win the business, you're ready to start building a proposal.

Last year you landed a similar deal, and you want to reuse that great work. You ask Copilot to use that proposal as a starting point and update it to include the new customer requirements. With one last click, Copilot pulls it all together for you.

This is truly a new pattern of work, bringing together content from the web and work in a dynamic, persistent canvas designed for human-to-human-to-AI collaboration.

Back to you. Satya.

SATYA NADELLA: Thank you so much, Callie.

It's fantastic to see this next phase of Copilot Plus Agent design system come to life, where what we started just 18 months ago with chat is now much richer in how we work, and it'll become natural part of, I think, everyday workflows.

I now want to talk about the Copilot Plus Agent stack. Everything that we used to build out, the products that you just saw, are also exposed as a first class platform for every software developer out there. That's really what the Copilot Plus Agent stack looks like, starting from the raw infrastructure to data to the AI tools and the AI app server.

Now, of course, one of the foundations of all this is these intelligent factories. They happen to be data centers. Microsoft has a comprehensive footprint of data centers. We have 60-plus Azure regions around the world, 300 plus data centers, and now, of course, as of May, we have a fantastic AI region right here in Mexico.

Today, I am really thrilled to announce that we are doubling down on bringing more capacity to Mexico by investing in additional billion.

(Applause.)

To us, this is at the foundation. Whenever we think about any region, bringing these data centers, these AI Azure regions to a market is what allows us to bring about a deep transformation in the local ecosystem's ability to take advantage of all of this new commodity as just an input, so that they can create value add.

It starts with the infrastructure layer. When you think about AI infrastructure, the core, the innovation at the silicon level coming from our partners, such as NVIDIA or AMD – we are building our own silicon – we want to make sure you have the choice of the best silicon that is powering this AI infrastructure.

On top of that, we have the Azure AI platform, which then allows you to be able to build your own agents, and that's essentially the same technology that we use to build our own copilots.

Now on top of that, you have the best selection of models. OpenAI has done leading work here. In fact, the new o1 model sets absolute new benchmarks when it comes to

reasoning. 4o was the best in class model from an LLM perspective and a multimodal LLM perspective. Therefore, OpenAI continues to lead the market, but on top of that, we also have a broad choice of models, whether it's open source models other closed source models.

All of these are available to you as part of Azure AI so that you can use what is the best model for your use case, because you always want to optimize for your cards, your latency, your performance on a specific eval. And so, you will have always the open system for all of the models that you can then use.

Now, it's not just the AI infrastructure and models that are changing. The entire data estate is fundamentally going through a massive transformation. Think about it, right, which is you want to be able to bring this new intelligence, this new reasoning, new planning capability. You want to ground these AIs with memory, context. That means data, whether it's the classic relational data, whether it is the NoSQL data, whether it is vector databases, all of these have to be transformed to be able to now work alongside your AI to create the next generation of applications.

I think some of the most exciting work that's happening even inside of Microsoft is in building out a new AI-aware OLTP stack, a new AI-aware analytical stack, which then every software developer can use in order to be able to build the next generation of applications.

Now around this is what is a rich tool chain and an app server. Essentially, just like how, back in the day, we built an app server around .NET, we are building now a rich set of tools and an app server around AI, so that you can do these model selections. You can upgrade from one model to the other. You can do these evals. You can do safety. You can do guardrails to ensure that you have the right classifiers around your applications, so all of that, plus the best tools for productivity.

GitHub Copilot went from a concept to a standard issue for every software developer, unlike any other diffusion I've seen in my life. It's unbelievable. I was able to sort of even see people right here in Mexico now, who use GitHub copilot every day. In fact, Mexico happens to be the second largest GitHub install base, with 1.8 million developers right here in Latin America. It's fantastic to see.

(Applause.)

And so, we are very, very excited about how software developers, organizations here are taking advantage of this Copilot+ agent stack.

I had a chance to meet with APEC, which is working on really bringing early detection to retinopathy and any retinal diseases, which is mission critical work, because early detection can make a huge difference there. They're able to use some new models. In fact, it's very nice, the way they've done it. Instead of trying to just take static images, they just record a video and then pick the frames, or in other words, AI picks the frame to

be able to then do the diagnosis. And that's changing, effectively, the early detection rates.

I already talked about what Cemex is doing across the length and breadth of their digital estate.

I also had a chance to meet with the folks from Monterey Tech earlier today to just see how they're enabling across the entire ecosystem, how they are implementing AI. Let's roll the video to get a sense for it.

(Video segment.)

(Applause.)

SATYA NADELLA: It's always fantastic to see how the rate of diffusion of any new technology just gets faster. When I used to come to Mexico, where in the early days of the cloud, you already saw cloud getting deployed broadly in the country, but when it comes to AI, I don't think that there's much of a difference, quite frankly, between what's happening, let's say, in the United States and Mexico. And that's fantastic to see the rate of progress of even the diffusion of anything new. And you can see that with the Copilot+ agent stack.

Now, I want to talk about the last platform, which is Copilot+ agent devices, obviously starting with the Copilot PC. In fact, very much like what's happening in the data center, you see that same revolution, that same transformation starting to happen on the edge devices, everything from the silicon innovation.

Right now, when you think about the PCs that are getting built with these new NPUs, starting with what's coming from Qualcomm, what's happening with AMD, what's coming from Intel, and there's going to be so much silicon innovation that's going to be back in the PC ecosystem that's so exciting.

And then on top of that, we're building a very rich runtime for hybrid AI. This is not about just edge AI on its own. It's the ability to use all of the flops that are there locally, but always have access to the (card?). No application wants to be less than perfect. And so, our ability to be able to have third parties write applications that exploit both local NPU capacity, but also tap into the cloud is going to be built foundationally into the developer platform that comes with these Copilot+ PCs, and then all the experiences, starting with the copilot itself.

We are very excited to see what even software developers are already doing with some of the models, both local and cloud, and then using, in fact, the PC as sensors. After all, you can think of it as you give AI eyes and ears and input, and so, the ability to sense the real world. And then to be able to use that with these rich AI models, both locally as well as in the cloud, I think, are going to be very transformative.

Let’s give you a bit of a feel For what the Copilot+ PC era looks like.

(Video segment.)

(Applause.)

SATYA NADELLA: It’s really exciting to see all of the advances. In fact, the other thing that I’m most excited about is even the fundamentals. I’ve been using now a Copilot+ PC for maybe the last five, six months. Just even basics, like battery life, it’s fantastic to see how what a difference it makes, even in the age of AI.

In fact, it’s one of the most exciting things for me is to have these applications now that are written that take advantage of the NPU. And it leaves the CPU and the GPU alone for other things. And so, therefore you can play your video game and have AI running simultaneously by being battery efficient. I think that type of system performance, I think, is going to be something that lot of software developers everywhere are going to exploit.

I want to end, though, by talking about, perhaps, what is the most important piece of work that all of us collectively have to do, but it starts with us, which is trustworthy AI around all these three platforms. Ultimately, you need, in digital technology, trust. When you think about trust, it’s all about consistency over time, across the commitments we make and the capabilities we build.

The commitments we make fundamentally starts with security. We initiated something called Secure Future Initiative in November. We are hard at work. In fact, just this week, we announced a progress report on what we are doing on making sure that we, as a company, take security above all else, whether it’s protecting identities, protecting networks, protecting tenants, being able to in a very fast rate, respond to any attacks. That’s what is the cornerstone of our work in security.

We have a set of very well established privacy commitments and principles that we adhere to. One of the fundamental things, for example, in the age of AI, is to ensure that your data is your data. Then when you use these models, and we don’t use your data to train those models, but you can use the models to be able to reason over your data and be assured that there is no leakage of information. Privacy is super important.

And responsible AI. Ultimately, you want to be able to deploy these AI systems, which are increasing – with increasing levels of sophistication, but know that you can have guardrails around them. You can ensure that there is no bias in them. It’s fair.

Those are all the things that we are building in and our commitments. And, of course, today, we are even excited to see some of the fantastic announcements on progress we’re making, when it comes to grounded-ness, for example, so that around safety, there’s some new features in Azure AI. We’re announcing new capabilities in confidential computing for privacy, new capabilities in security.

Overall, we continue to advance our capabilities, but always focused on trustworthy AI and the commitments around security, privacy and safety. And so, we are very, very excited about this. Ultimately, all this aggregates up to our mission, which is to empower every person and every organization on the planet to achieve more. And it starts with what we're doing in Mexico.

To me, ultimately, technology being able to make a real difference to a society and to a country and its economic prospects and economic growth, I think, is what really brings home our mission. That's what excites everybody at Microsoft.

And one of the things that I think about in the age of AI is as technology is important, ultimately, it's the human capital of a country that uses this technology to translate it into real-world impact. It's better health outcomes, better outcomes in industry, better outcomes in public sector, and it always starts with human capital.

We are very excited to announce that we will be skilling 5+ million people in Mexico around AI. And so, this is an exciting posture.

(Applause.)

Thank you very much for all that you do. I'm looking forward to seeing how all these three platforms really get used intensely in Mexico to create much greater technology coming out of Mexico that not only has impact here, but around the world.

Thank you all very much.

(Applause.)

END