05212024 Build Rajesh Jha

Microsoft Build 2024
Rajesh Jha
Tuesday, May 21, 2024

ANNOUNCER: Please welcome Executive Vice President, Experiences Plus Devices, Rajesh Jha.

RAJESH JHA: Good morning. I'm Rajesh. You heard Satya talk about the Copilot Stack, the AI architecture of the future. Now, I want to talk about bringing that AI stack to Microsoft products. First, I'm going to focus on how we are expanding Copilot. Then, Jeff Teper will talk more about extensibility, and then finally, Pavan Davuluri will share more on Copilot+ PCs and the Windows ecosystem.

Now, diving into how Copilot is evolving, it was only a year ago at Build that we showed you the promise of AI, and then we made Copilot for Microsoft 365 generally available in November. I want to spend a moment and talk about the journey that we've been on. Within Microsoft 365, we brought Copilot to the applications that hundreds of millions of people use every day. Giving our users a powerful new way to interact with AI, right in the flow of their work.

Now, we've seen great adoption across a range of customers and industries. Nearly 60% of the Fortune 500 now use Copilot. And we've seen accelerated adoption across industries and geographies with companies like Amgen, and BP, Cognizant, Moody's, Novo Nordisk, Nvidia, Tech Mahindra, and many others are choosing over 10,000 seats each. We've added over 150 Copilot capabilities since the start of just this year.

In addition, we continue to integrate Copilot across more of our productivity apps and services, from OneNote to Stream, to Forms, to OneDrive and more. Now, we've also launched a standalone Copilot application. Whereas Copilot had been embedded in Office, Outlook and Teams, the Copilot app now has all of Microsoft 365 embedded in it. And what makes the Copilot app unique and uniquely powerful is the grounding it has.

First, our Copilot understands the web. More than the grounding the Copilot is crucially grounded in the user's work context. Who do they work with? What do they work on, their meetings, their conversations, the documents? And that is the Microsoft Graph. The Microsoft Graph represents the user, their teams, their permissions, their organization. It represents their context.

Now, it's important to note that the Copilot app is no different from any of our other commercial services in terms of compliance and data handling. Microsoft has no eyes on access, and your data is not used to train the models. So, let's dive in and see how all this comes together in the Copilot app. In the web tab, you can get answers to simple or complex questions grounded in live web data, all with commercial data production.

Now, let's toggle over to the Work tab. Here, Copilot is grounded in Microsoft Graph, meaning it has access to your personalized work environment, the people you interact with, important files, teams, meetings, all your communications. By the way, this is not a demo account. This is my actual Microsoft account, using real data to show you how personal this experience really is. Of course, you're going to see some redactions; I do want to keep my job.

But learning to ask the right question is key, so you can get the most of the Copilot. To help you, there's a prompt library to give you suggestions just when you need them. Let me start by getting caught up on the latest from my boss, you may have heard of him. Copilot understands organizational structure, so it knows Satya is my boss. It scans the latest emails and chats and files, and in moments, I have a detailed breakdown with updates that need my attention.

Let's try one more. Here, I am asking Copilot to propose a session title based on this very keynote script, which is a document in SharePoint, and to create a list of topics on Generative AI to discuss in this session. Copilot has analyzed a document from my work environment and proposed a suitable headline. But since Copilot is also grounded in web data, it is able to recommend topics sourced from the web for me to consider.

Now, switching gears. Since its inception, Copilot has been a uniquely personal assistant, as you've just seen, but we all work in teams, small and large, organizations, intimate and global, and we want to do more to go beyond, to empower people when they come together. And like Satya said, today we're announcing Team Copilot, the expansion of Copilot beyond a personal assistant. This will enable Copilot to serve and act on behalf of a team, a department, an entire organization, not just an individual user. Copilot will act as a valuable team member, improving collaboration, project management. Let's take a look.

(Video segment)

RAJESH JHA: A Team Copilot will be a valuable new member of any team, and these initial capabilities will be available to our customers in preview later this year. Now, in my conversations with customers, one of the top questions I get is, "How do I translate productivity gains into transformative business results?" And simply put, the answer is moving to a reimagining of business processes using their own Copilots and agents and extending Microsoft Copilot.

Now, Jeff is going to walk you through how our extensibility to platform is going to enable you to do just that. But first, let's take a look at the Copilot architecture that really makes all of this possible. A Copilot, as we spoke, can recall and reason our up-to-date web knowledge. It is grounded in Search. All Copilot experiences have this web scale.

Now, in addition to being grounded in the web, it is also grounded in your data through the Microsoft Graph. And when the Copilot is in an application like Office or Teams or Edge, it also understands the application context. So for example, in PowerPoint it is able to draft a slide deck for you from a document.

The Microsoft Copilot is architected to compose or inherit capabilities based on the user context. And now, as a developer, you can build Copilot extensions at the data layer, at the experience layer, to further extend and customize the Copilot. And all of this is enterprise grade, with tools for IT to manage and personalize for employees. Our leading ITs are already working with us, building solutions that can extend the Microsoft Copilot, and I would like to share two examples from ServiceNow and Adobe for ServiceNow.

As you know, ServiceNow helps organizations orchestrate and automate tasks and processes across their enterprise. Here, we see ServiceNow's Copilot Extension now assist responding to user prompts inside Copilot for Microsoft 365 with the exact same knowledge, functionality and user experience as it has today in Teams. ServiceNow has included several custom zero query default prompts to help users get started with the most common tasks, without having to know how to crack the right text to start the conversation.

Over to Adobe. For working on bringing Adobe Experience Cloud workflows and insights to Microsoft 365 in Copilot, with Adobe Express Copilot Extension, users stay in the flow of their work in a Word document and can start a workflow in Adobe Express that allows them to create social content, select and edit images and stages for publishing.

So, as we close out this first chapter, I hope you're excited. Microsoft Copilot is already helping people save time, be more productive and creative. Team Copilot expands Copilot in meaningful ways, and then there are great developer opportunities for you to extend Microsoft Copilot. Jeff Teper will join us now to share how easily you can build Copilot Extensions.

But first, I want to close my session with a video showcasing how Lumen is using Copilot for personal productivity, to enhance their sales processes and connecting Copilot to their system using Copilot Connectors. Let's roll the video.

(Video segment)

JEFF TEPER: Well, as Rajesh shared, you will be able to easily and securely use your applications and knowledge to build Copilots that help your employees and organization be far more productive and grow your business. And you can now extend the Microsoft Copilot with your own Copilot with handoffs in all the Copilot experiences, as well as in Microsoft Teams, where you can reach hundreds of millions of users today, for both the personal and group assistance scenarios that Rajesh outlined.

And we are making building these Copilots even easier. From a few clicks in SharePoint to more advanced customization in Copilot Studio, to full control of your models, your data, your applications, your actions, your experience in Visual Studio Code.

Let's first look at what this means for end users. Copilot Extensions run everywhere that Copilot is, the standalone experience across Microsoft Teams, and as we're showing here in the Microsoft 365 app. On the right, you can easily browse your installed Copilot Extensions, find new ones, or build your own, which I'll show coming up.

How Microsoft Copilot works is it reasons over the user's prompt and maps it to the right extensions. Or you can explicitly app-mention that extension like we're showing here. You're going to be able to drill into a deeper, focused conversation with the extension, like we're doing in this marketing example, that has suggested prompts for quick actions and to just show the users the capability of your Copilot. This allows the Microsoft Copilot to have real-time access to knowledge and applications in your environment.

Here, we're using a suggested prompt to ask about a key feature of a delivery drone. The Copilot extension you build is going to come back with a visual adaptive card, bringing in all the information to avoid an unnecessary multi-turn conversation so the user can just focus on getting their work done. And again, these Copilot extensions also run in Teams, in one-on-one, in group chats, and channels, and in meetings, so you can reach all these users today.

All right, this is Build, so let's get to building with our first custom Copilot extension. You're going to be able to do this from again, a few clicks in SharePoint to advanced customization in Copilot Studio, to Visual Studio Code. We'll start in SharePoint, which is often the authoritative source of knowledge and content processes with advanced collaboration, workflow and security, all of which Copilot honors to make sure users only get access to information that they have permission to.

I'm in the SharePoint site. I'm going to go ahead and select a few documents, hit "Click to Copilot," and right there that looks good. I'll go ahead and change the name of this to the delivery drone. That looks fine. And just like that, I've created my first custom Copilot that you can use to extend the Microsoft Copilot.

Let's go ahead and try this one out. How much does a delivery drone service cost? And we can see it comes back with a flat fee of $5 per order. All that looks pretty good, secure, grounded. We're going to go ahead and share that with our team, and up comes the standard sharing dialog to honor the security in your organization. We're going to copy this link, go into Teams, paste it in a chat that will say, "Try this out." We will go ahead and paste that, and just like that, in seconds we've created a secure, grounded, custom Copilot and shared it with our team in Microsoft Teams.

We're very excited about letting anybody create these secure custom Copilots, and this support will be available in SharePoint this summer. Sign up today for the preview. We're very excited about that.

All right. Next on the spectrum, we're going to do some more advanced customization in Copilot Studio. Starting right from SharePoint, I can launch into Copilot Studio for my more advanced edits, and you can see all the information for the Copilot I just created is carried forward, so that's all there from SharePoint. Let me test this out by asking when the launch event is, and it'll return back that it's on June 20th.

But what I really want is that Copilot to do work for me, not just answer questions. And so for this, I need to go into Copilot Studio and start by adding additional data sources in the knowledge tab. Here we can add websites and files and connect to over a thousand Copilot

connectors. In this case, our account information is in two tables in Dataverse, and so I'll go ahead and select them, and we've got the information we need.

Next, we move to the Actions tab, where I need to see if somebody's already registered for the event and if not, send them a personalized invitation. Our event registration is managed in an external system, so we've created a custom connector to go get it, and you can see us configure that. We can tailor the action and input and outputs however we need.

Then the second step is to automate the sending of that personalized invitation. We've built a custom power automate flow, so we'll add that as a Copilot action, as well. And there you go, and pretty quickly we've got a complex Copilot extension with content from SharePoint, data from Dataverse, two disparate actions, and we can go ahead and test this in Copilot Studio to see if Contoso is actually attending the launch event.

It says they're not, and the Copilot conversation helps you see why what's happening, and map that all through. We can ask a follow-up question, again that's routed to Dataverse, about who the account manager is and we can see that's Perry Lang. Then we can ask Copilot to go ahead and send the invitation.

Now again, what's happened here is the conversation was entirely generated for me, identifying and chaining together the key knowledge from Dataverse and the appropriate actions we added earlier with Generative AI capability.

Last, we're ready to publish this Copilot extension back to SharePoint, Microsoft Copilot and Teams, and from Teams you can search for it in the unified marketplace. I can add it to one-on-one or group chats or meetings, etc. And again, I can see this in the Microsoft Copilot experience, that same relic cloud Copilot and continue the experience there.

So, we're very excited. This is a pattern that people are already doing today. Let me show you what one of our customers, Wolters Kluwer, who is a leader in information and software and solutions, is doing to enhance their tax and accounting professionals' workflow with Copilot. They're building a Copilot extension to enable their accountants to complete each step of their workflows, interacting with their backend system, just using natural language with no context switching.

Copilot takes actions on their behalf, saving time and cutting the process down from minutes to seconds. And when it's time to communicate back to their client, the extension helps close the loop, drafts an email, attaches the estimates and ensures a seamless end-to-end productive experience.

We're excited about all of you building these kinds of custom Copilots and Copilot extensions. Copilot Studio is now generally available to build enterprise-grade Copilots. The new capability to publish Copilot extensions from your Copilot is in private preview. And as you heard Satya say earlier, Copilot Connectors are now in public preview and they make it even easier to connect your Copilots to your business data, your apps and workflows.

OK. Last, we want to show you how to build a Copilot extension as a professional developer with full control of your models, your data, your actions, your experience, in Visual Studio Code. So, let me go ahead and do that.

What we've got here is Visual Studio Code with the Teams AI toolkit installed, and we've loaded up a template for building a custom Copilot and using it as an extension. The first thing you see in the code is where I configure the model. You can use an off-the-shelf model. Here we're using one from OpenAI. You can use a refined model, or you can build your own completely tailored to what you need. If we go scroll down the code a little bit farther, this is where RAG integration is. You can see in a few lines of code we've integrated our data; in this case Azure's vector search capabilities. You can use any other data source; the Azure AI search capability is a great one to use. We keep scrolling down, we see the actions defined and registered.

But here is where we go search for our product inventory, and you can see we've inserted a breakpoint so that we can follow this along because I want to show how easy it is to do end-to-end development here. We'll bring up Teams where we've installed that Copilot extension for the inventory, find information for the Chai tea product, hit return, and we're paused. Why? Because we've hit that breakpoint. We can go back into Visual Studio. You can see sure enough it's fired. And if I hover over the parameters, you can see the product named "Chai" has been passed to it. This is incredibly powerful right from within Visual Studio to do end-to-end debugging across the Microsoft apps, Copilot and your Copilot extension.

If we keep going down before we return that to the user, I just wanted to show what the UX looks like. Here's an adaptive card that comes back. It's defined in JSON, but you can also see the user experience for that. That all looks good. Let's go ahead and resume execution from the debugger. You can see that we've now returned back into Teams. The result, and we've got an attractive adaptive card that gets the user the information they need, does hand off with the Copilot, so right in line, the user can complete the work. Pretty exciting. Again, full flexibility within the Teams AI library and Visual Studio Code. We're excited, just like with Copilot Studio, this is something you can do today targeting hundreds of millions of Teams users. And very soon, the ability to turn your custom Copilot into an extension will be available.

This is not new, this is something that hundreds of ISVs are already doing today across all sorts of experiences in Microsoft 365, Teams and now Copilot. Just two examples are leading software organizations Esri and Thomson Reuters. Esri is the market leader in geographic information systems. They're building a Copilot extension that adds spatial analytical capabilities directly into Teams meetings so that users can ask Copilot for Microsoft 365 a question about map data, and Copilot will seamlessly hand off to a rich, interactive experience with visualization in Esri's custom Copilot with all the associated context. Next up is Thomson Reuters, who's a leading global end content and technology company that is transforming the legal profession with AI. Thomson Reuters is extending the Copilot experience in Outlook and Word in Teams for things like risk assessment, so that based on the content of the Outlook email, they can update the policy documents in Word and communicate those policy changes to reduce risk right within the Teams meeting.

We're super excited to see what you do targeting this huge user base, and we're going to help promote and distribute your application through our unified marketplace, again, reaching hundreds of millions of users today in Microsoft 365 and in Teams. One of the reasons this marketplace is trusted is IT has confidence in the full governance capabilities for their own applications and custom extensions they build, as well as the ones they will get from all of you building them in a vibrant third-party ecosystem.

To recap, we have a simple, powerful platform for AI in Microsoft 365 that you can use to be far more productive across the full spectrum from something everyone can do in SharePoint to advance customization in Copilot Studio, to the full power of Visual Studio and Visual Studio Code. But wait, there's one more thing. We are very excited to make Teams a fantastic place for developers to work together with AI to write better code faster, so we've got a whole set of announcements around that this week at Build as well. First things first, source code inside Teams with syntax formatting. And get this -- we wondered when we'd get the applause, and that was my bet. But wait, you should have held the applause. With Microsoft Loop, co-editing of that source code right within Teams. But there's more.

Developers have asked us for a while for greater information density in Teams so you can create and switch to compact mode to see much more content on the screen, and you can be much more productive with things like keyboard shortcuts and new slash commands for Teams. Of course, developers are always in the flow of resolving issues in chat. One of the features we're really excited to announce is "Meet Now," so that right within chat, you can bring up a ringless call between members of the team and resolve the issue in seconds. And last, developer teams love to have fun to break from the stress, so you can use custom emojis with reactions now in Microsoft Teams.

Of course, this is building on top of a growing set of partnerships with DevOps tools. Jira, Datadog, PagerDuty, and of course, deeper integration with GitHub and much more integrated with Microsoft Teams. We are very excited to make Teams a great place for developers to work together to build this next generation AI. The next chapter is how Windows is the best platform for building that next generation AI. And to show that, I'm excited to invite Pavan to the stage. Pavan.

PAVAN DAVULURI: Thank you, Jeff. Good morning. It is great to be here at Build. This is one of my favorite times of the year, connecting with fellow product makers about the world's canvas for innovation, Windows. I'm excited about the Copilot extensibility that Rajesh and Jeff just shared. It really shines on Windows; the platform customers choose for Microsoft 365 and Copilot. Over the last year, we've learned so much about how Copilot can best serve you. We're working hard to make it even more valuable with a vision for Copilot meeting you right in your workflow. Imagine creating a presentation from a document in File Explorer or helping customers troubleshoot their PCs using quick actions and natural language right in settings. We're focused on making Copilot even more contextual and useful across Windows. We took a big step towards that goal yesterday with the announcement of Copilot+ PC, the fastest and most intelligent PCs ever built. Let's take a look.

(Video segment)

That is inspiring. AI is woven into every layer of these devices, from the silicon to the operating system, with the most powerful PC NPUs capable of delivering over 40 trillion operations per second. This new class of PCs is up to 20 times as powerful, and 100 times as efficient for running AI workloads, compared to traditional PCs from just a few years ago. Built together with our silicon partners AMD, Intel and Qualcomm and our OEM partners, these PCs will be available June 18th, starting with Qualcomm's Snapdragon X series of chips. Copilot+ PCs are redefining what you can do on a PC and setting the direction for the next decade of Windows. To put this new wave of AI innovation in your hands, we're excited that Qualcomm has announced Snapdragon Dev Kit for Windows. It is designed to be your everyday dev box for AI with the power and flexibility you need.

As we define this new path for Windows in the era of AI, one thing that will never change is our commitment to openness. We recognize that the real value of Windows comes from the energy and the innovation of the ecosystem. It comes from all of you. As we enter this new era, let's talk about how we're going to serve over a billion Windows customers together. As Satya said earlier, building a powerful AI platform takes more than a chip or a model. It takes reimagining the entire system from top to bottom. The new Windows Copilot runtime is the system that extends the Copilot stack to Windows.

The Windows Copilot Runtime is a new integral part of Windows 11 and has everything you need to build great AI experiences whether you're just getting started or already have models of your own. It includes the Windows Copilot Library, a set of APIs that are powered by on-device models that ship with Windows and includes AI frameworks and tool chains to help you with your own on-device models. It's built on the foundation of powerful client silicon, including the NPUs in the Copilot+ devices. Let's take a look at how the Windows Copilot Runtime enables an entirely new class of experiences. OS experiences like Recall that help users find anything they've seen on their PC, in-box app experiences in Photos and Paint, which let you bring your ideas to life using real-time image generation, and app experiences like CapCut, Cephable and DaVinci Resolve, some of our first partners using the new NPU and helping us build the Windows Copilot Runtime. Looking ahead, the Xbox team has a vision for using the Windows Copilot Runtime to empower players and game developers. Let's take a look.

(Video segment)

That's pretty inspiring. This entire class of new experiences now benefit from faster task completion, enhanced privacy and lower costs by using the Windows Copilot Runtime. Next, let's take a look at the Windows Copilot Library, the APIs and models that support them. Let's take the Recall experience as an example. It relies on on-device models deeply integrated into Windows to capture context on the screen. That data is transformed into vector embeddings and index into vector store, called the Windows Semantic Index. The Recall User Activity API allows you to extend your app into Recall so users can jump right back to where they were in your app and increase your app engagement in the same way Edge and Microsoft 365 apps like Outlook, PowerPoint and Teams already have. In fact, soon, Recall will draw in context from the Microsoft 365 graph.

To build your own semantic index store, you can use the Vector Embeddings API. That makes it possible to use retrieval augmented generation, RAG, within your applications with your data. Imagine you have a WinForms or a WPF app that works against a large corpus of sensitive data. With Vector Embeddings API, you'll be able to create on-device vector stores for those records. That's powerful when combined with the RAG API to enable natural language search in your applications for your users. Of course, that's just one example. The APIs in Windows Copilot Library cover the spectrum from low-code APIs to sophisticated pipelines, to fully multimodal models, like the recently released Phi-3, the single best SLM in the world.

Phi-3 mini does a better job than models twice its size on key benchmarks. Today, we're thrilled to announce Phi Silica. Built from the Phi series of models, specifically designed for the NPUs and Copilot+ PCs. It offers lightning-fast, on-device inferencing and state-of-the-art first token responsiveness. Windows is the first platform to have a state-of-the-art SLM custom built for the NPU shipping in box. Now, let's take a look at what you can do to bring your own on-device models to Windows using frameworks and toolchains. It starts with DirectML, the lowest level machine learning framework in Windows, similar to DirectX for graphics.

Whether it's your own open-source models or an open-source model from Hugging Face, DirectML helps you scale the breadth of your efforts across the Windows ecosystem by giving you to-the-metal access to GPUs and NPUs. We also know that a lot of you do your development on PyTorch on Windows, and we're thrilled to announce that Windows will natively support PyTorch through DirectML. That's right. Pretty exciting. Native PyTorch support, of course, means that Hugging Face models will just work on Windows, and not just that, we're collaborating with Nvidia to bring these workflows to over 100 million RTX GPUs in the Windows ecosystem. Now, that's incredible. You can download the PyTorch and DirectML Developer Preview today.

We're also going to extend DirectML to our web developers by introducing WebNN on Windows. WebNN is a web-native machine learning framework. Microsoft has been working with Intel and other partners to unlock the access to local ML accelerators, so you can build performant AI experiences in your web apps. Behind me, you see Clipchamp's Auto Compose feature, achieving faster video composition experiences and cloud savings by leveraging the NPU for that and WebNN. I'm excited to announce that WebNN is available in Developer Preview today.

That's a glimpse of the Windows Copilot Runtime and how it lays the foundation for innovation, giving you the largest catalog of models on the largest ecosystem of devices, making Windows the most open platform for AI. I heard you there; that's fantastic. As Windows transforms for the era of AI, we're continuing to reach the expanse of the platform, including all the AI experiences you create with the Windows Copilot Runtime. We're delivering Windows from the cloud with Windows 365 so your apps can reach any device anywhere, and we're introducing Windows experiences to new form factors beyond the PC.

For example, we're deepening our partnership with Meta to make Windows a first-class experience on Quest devices. Windows can take advantage of Quest's unique capabilities to extend Windows apps into 3D space. We call these "volumetric apps." Let's take a look.

VIDEO DEMO: Workflows are transforming with mixed reality. Microsoft is partnering with Meta to bring Windows 365 and local PC connectivity to Quest and enable developers to easily extend their Windows apps into the 3D space. PTC has been working with this platform, bringing Creo into mixed reality in under a day. This extension allows users to enhance spatial understanding without leaving the app that powers their work. Sign up for the Developer Preview today.

PAVAN DAVULURI: That's good to hear. As developers, you'll have access to volumetric API, and this is just one of many ways to broaden your reach through the Windows ecosystem. For decades, Windows has been the stage for the world's innovation. With Copilot+ PCs, the Windows Copilot Runtime and Windows 365, we're going to unlock a new era of innovation together. Thank you. Back to you, Rajesh.

RAJESH JHA: Thank you, Pavan, and thank you, Jeff. We've covered a lot of ground over the last 40 minutes. From the expansion of Copilot beyond a personal assistant to acting as a valuable team member, to how you, as developers, can extend Microsoft Copilot with your own Copilots and agents in just a few clicks in SharePoint to more advanced customization in Copilot Studio. You can use VS Code full control of your models, your data and actions. Of course, a phenomenal opportunity for developers with over a billion Windows customers. I'm going to close with highlighting another customer, Amgen, a pioneering biotechnology company that harnesses the power of biology and technology to fight the world's toughest diseases. We are going to see how they harness Microsoft Copilot in their mission. Kevin Scott, our CTO, along with some special guests, will round out the day one of keynotes. But before the video, let me just finish by simply saying thank you. Thank you for spending your time with us here at Build. It means a great deal to all of us. Thank you for the trust you place working together, building the future with us every single day. Let's roll the video. Thank you.

(Video segment.)

END