David: Welcome to the Microsoft industry experiences team podcast. I'm your host, David Starr, and in this series you will hear from leaders across various industries discussing the impact of digital disruption and innovation, sharing how they've used Azure to transform their business. You can find our team online at AKA.MS/INDXP or on Twitter @industryXP.

Phani Mutyalais is a senior program manager with the Applied AI team at Microsoft. He focuses on running container strategy and execution for cognitive services. Phani has a deep experience in building and running large scale platforms on public clouds and, more recently, with container platforms. He's super passionate about cloud, containers, machine learning, and AI technologies and is a continuous learner. Welcome to the show, Phani.

Phani: Thanks, David. Really excited to be part of the show.

David: Yeah, we're glad to have you here. Maybe to start off could you tell us a little bit about what Microsoft's Azure Cognitive Services are?

Phani: Sure. So today machine learning in general requires machine learning experts or being a data scientist to dial up the frameworks of the models. So cognitive services in general, they are pre-built, which are developed by Microsoft with our own data. These machine learning models which are pre-built as specialized pieces of AI, which are meant ready to use without being a data scientist. So these APIs are easy to implement because of the simplest rest API calls. We got a breadth of intelligence in this APIs dollar per any person would be able to find what intelligence feature in this API today and what they need. And, most importantly, they all work on whatever language or framework or platform that our developers choose. Any depth they can integrate into our cognitive services APA model into the app such as IOS, Android, Windows using their own tools they know and they really love. Such as Python, [inaudible 00:02:23], etc.

So, developers can trust the quality and expertise built into this API cognitive services which are built by our experts from Microsoft Research, and these capabilities are used across many Microsoft [inaudible 00:02:37], Bing, and Skype.

David: That is a big set of capabilities coming from Azure Cognitive Services. What is the idea behind containers for cognitive services? What problem does that solve?

Phani: Sure, so today these cognitive services API really being allowed by a lot of industries, like different verticals or different industries. But there are some industries only interested in what they can do with the AI, you know? That is all because of AA services we have built are running in the cloud. And with a few customers and a few enterprises, because of the regulatory reasons such as [inaudible 00:03:17] or other reasons maybe due to data governance or performance, they're unable to actually call those services in the cloud. So, containers from cognitive services open up a whole new world of possibilities allowing the customers, enabling them to take that at one age at the full power of AI right in their own control, in their own network. Basically, on their own terms, like how

they want to use it, definitely scale up, scale down. The services completely in their own control. That's the beauty of containers from cognitive services.

There are a couple of other reasons, also. In a few cases, customers are not willing to share the data or load the data into the cloud. For processing, not for storage. And somewhat due to [inaudible 00:04:02] handling the sensitive key of their customer data. And these are the few reasons why they care about the containers to run completely locally on premises, not on the edge.

David: For people who haven't heard that term before, what does "on the edge" mean?

Phani: So, edge GYS could be anywhere, right? It could be on customer premises or could be anywhere or everywhere in the world. So, customers completely run off the data centers. For example, the GYS edge could be a small robot or GYS could be a camera. It could be a recorder, anything that is portable is an edge GYS, or could be a bigger one could also be an edge GYS, so it's completely off the premises. Let's say a retail store wants to run the face recognition or detection on one of their camera, which can be considered as edge GYS. So, that's what edge GYS meant.

David: Okay, perfect. And I'm wondering now about real life, right? So, let's try to apply this to a couple of real industry problems and what might we solve in the real world. So, let's start with, say, retail. What can we do with containers from cognitive services in retail? What's that gonna buy us?

Phani: Sure, in retail I want to process customer feedback about products that I have in the catalog. I could run them through the language deduction and the key physics extraction being able to process the text with sentiment models help me to understand what's being said is positive or negative and how should I improve my product selection in the store. So that's one of the use case I can use the AA models running closely on my XTY [inaudible 00:05:39] could be locally within my store.

David: Take another one. How about manufacturing?

Phani: So, manufacturing is really interesting. We have a couple of customers share their feedback, and let's say, for example, one of the remote oil rig may not have a proper bandwidth because of the location that they are. And they want to transfer all of its data back to the cloud, which is very difficult. So they want to run the AA models closely within the oil rig because they have very strong needs to run it locally.

David: In that model I assume you would come along and a technician might come along and pick up the results of the learnings over time off of an SD card or something.

Phani: Exactly. Basically on edge GYS they can run ... they can take the model and adopt a container and they can deploy it on the edge GYS on the premises within the oil rig environment and they can get the results then and there itself without being every time talking to the cloud.

David: That brings up a good point. How does this impact the architectures that developers might be building today if they were to move to this model of containers for cognitive services? What's going to be different about their applications?

Phani: So, it's not going to impact anything. So it's up to the developers to use cognitive services system within their own control of the data. For example, so far developers let the business application talk to the [inaudible 00:07:04] endpoint. In this case, they talk, they let the business application talk to the container endpoint. It support the consistency in the hybrid environments, like in across the data, management, identity, and security. So definitely the developer has more control over the data with all these functions. And also not just that. Also effects the flexibility in [inaudible 00:07:27] in building of the models deployed in the solutions. It also enables a creation of a portable application architecture that they can deploy anywhere and everywhere as they like. Definitely giving more control to the developer than earlier.

David: A device might be out there on the edge, and it's running cognitive services in its own run time. Is that correct?

Phani: Yes, that is correct. So, we deployed the cognitive services model within that XGY. It has its own run time. So, based on the configuration we assigned to the container, it will run. It will generate output as they want.

David: And what cognitive services are available now?

Phani: Great question. So we recently announced a few containers from cognitive services. So we have face detection and face recognition. We have a fitness of text, formally called OCR from vision services. And from text olympics family we have language detection, sentiment analysis, and key physics extractions. And also less than two weeks we released a new container which is called language understanding. So there are six containers available totally I public today for our customers.

David: One of the industries that I forgot to mention, but it's very interesting in this context is healthcare. Can you speak to some sort of use case for healthcare?

Phani: Sure. Let's say you own a hospital and have thousands of aggregte documents and you need to start processing those. So with condensed text in containers, you can actually run them all on [inaudible 00:09:02] feed on the documents. We thought sending them to the cloud, and absolutely then and there you get a distillation of all the information locally. So that's one of the use case most of our customers are being used.

David: Somebody's interested, how do they get started?

Phani: It's pretty easy. We don't have introduced any new sign-up process or completely, you know, separate process to sign up a container. So, you go to the [inaudible 00:09:29] portal, and you select the machine learning where you will find multiple models available. And let's say you are interested in face or you're interested in language detection. As soon as you create the resource it will give you two options: whether you

want to talk to the web poster in point or you want to download the container. And there are clearly instructions for doc container where you can follow download the image and run it locally, as we indicated in the [inaudible 00:09:57] dot com documents.

David: Well this has been great. Sounds like it really opens up a lot of opportunities for putting AI cognitive services in places where maybe we hadn't thought about before.

Phani: Absolutely. These are great opportunities for our developers to start running the cognitive containers locally because of a lot of challenges that they have been so far for the different reasons that we discussed earlier. This opens a whole new [inaudible 00:10:24] possibilities to run locally and on the edge to solve the problems.

David: What are some of the underlying technologies of the service itself?

Phani: So, these are X64 [inaudible 00:10:35] containers compatible to run with any orchestration, like [inaudible 00:10:41] Marathon or any kind of orchestration that can support the docket containers. And these are all CPU bound as of today.

David: Phani, this has been a really good information dump for me. I've learned a lot today, and I really appreciate you stoppin' by and talking with us about this.

Phani: Absolutely. My pleasure, and I'm super excited about customers taking that one page off, running the containers for cognitive services in their own environments according to the use cases that they have. We really excited to see how customers going to solve the problems running our containers. I'm super happy that you interviewed me. I'm really thankful for that. Thanks.

David: Thank you. Thank you for joining us for this episode of the Microsoft Industry Experiences Team podcast, the show that explores how industry experts are transforming businesses with Azure. Visit our team at aka.ms/indXP. And don't forget to join us for our next episode.