EPISODE 25

[INTRODUCTION]

[0:00:01.1] JM: A self-driving car needs to be able to quickly respond to changes in driving conditions. A factory needs to be able to quickly respond to changes in workplace safety. For these kinds of applications we need processing power closer to the user of the application. If we put all of our application logic in the cloud, we're going to have to make network round trip requests for every request that we send to the cloud. Servers in the cloud are powerful, but so are the computers at the edge; smartphones; sensors, drones, cars. You can even put servers on-prem, which is used to be the norm.

Today, edge computing is giving us more computation outside the data center. We're getting towards a world where you've got compute in the data center that you do some of your compute in and you've got computed the edge, that you do other types of computing at and we're going to talk about those different use cases in today's show.

Olivier Bloch works on Microsoft Azure IoT Edge, a set of services for edge computing. Azure IoT Edge includes on-prem versions of Microsoft Azure technologies. These are tools that were previously accessible only in the cloud and Azure IoT Edge allows you to deploy and host them on-premise.

Before we get to today's show, Software Engineering Daily is looking for sponsors for Q3. If your company has a product or service or if you're hiring, Software Engineering Daily reaches 23,000 engineers all listening daily. You could send me an email, jeff@softwareengineeringdaily.com if you've got something interesting and you're looking to sponsor the show. With that, let's get to this episode; IoT Edge with Olivier Bloch.

[SPONSOR MESSAGE]

[0:02:08.1] JM: Do you want the flexibility of a non-relational, key-value store, together with the query capabilities of SQL? Take a look at c-treeACE by FairCom. C-treeACE is a non-relational

key-value store that offers ACID transactions complemented by a full SQL engine. C-treeACE offers simultaneous access to the data through non-relational and relational APIs.

Company's use c-treeACE to process ACID transactions through non-relational APIs for extreme performance while using the SQL APIs to connect third part apps or query the data for reports or business intelligence. C-treeACE is platform and hardware-agnostic and it's capable of being embedded, deployed on premises, or in the cloud.

FairCom has been around for decades powering applications that most people use daily. Whether you are listening to NRP, shipping a package through UPS, paying for gas at the pump, or swiping your VISA card in Europe, FairCom is powering through your day. Software Engineering Daily listeners can download an evaluation version of c-treeACE for free by going to softwareengineeringdaily.com/faircom.

Thanks to FairCom c-treeACE for being a new sponsor of Software Engineering Daily, and you can go to softwareengineeringdaily.com/faircom to check it out and support the show.

[INTERVIEW]

[0:03:44.8] JM: Olivier Bloch is a program manager with the IoT Edge team on Microsoft Azure. Olivier, welcome to Software Engineering Daily.

[0:03:51.7] OB: Thank you, Jeff.

[0:03:52.4] JM: Over the past 10 years, we have seen a lot of computation shift from on-premise to the cloud, and over the next 10 years it looks like we're going to see even growth in cloud but also growth in devices at the edge, and today these are things like sensors in a factory or your internet connected refrigerator or your smartphone, and there's going to be growth in drones and self-driving cars and all of these things are computers and they have spare computation power. This means that we can push computation to the cloud or to these devices.

Generally speaking, when we have this huge array of different types of devices in the world as well as this cloud that is this giant area of computation, how do we decide where to put our computation?

[0:04:49.9] OB: It all depends and comes down to the type of scenario your implementing. Just to go back on what's just mentioned, in think that the evolution over the last 10 to 20 years on the device side of things has been more about actually having this device as being more more powerful and more more smart and more more connected, because the capabilities of the micro-controllers, these little tiny computers running there as well as the price of memory going down, what we are seeing is a huge growth in terms of what these devices could do.

At the time where it was not powerful enough to actually make decisions based on data or store enough data to do analytics on that data, obviously, the cloud has been here to the rescue to handle all these computers and storage tasks. Now what we're seeing is that's the device side of things has become such that it's not possible to start distributing your intelligence. It's not so much about doing it back-and-forth, it's actually about doing the right thing at the right place. Now, because the device is a more powerful, what you can start doing is not send all the data to a remote computer unit. You can actually now start doing things here locally on your device and you can specialize this tasks for what the device is doing.

I think when it comes to, "Hey, where should I run what?" It very much comes down to what is your scenario? What is it trying to achieve? Where are the critical aspects of your solution? This is where actually you enter into this kind of discussions. Typically, a question you will ask yourself is, "Do I need and do I want to pay all that bandwidth to send on that telemetry data to the cloud?" Because in a nominal situation, everything goes well, might not need that the. What I need is actually to receive information when something goes wrong, in which case you ask questions yourself which is, "Hey, can I determine that something is wrong down at the edge on the device or not?" This is where you start asking, "Hey, can I run this specific algorithm down close to the sensors or do I need to run it in the cloud based on how much data and how much power I have locally?" Then it comes down to that of scenario based context that you have.

[0:07:14.6] JM: I have done these shows with the Netflix. I talked to Netflix about building their scheduler and they build schedulers essentially entirely for cloud resources. Just, if you're

looking at a data center and you've got this number of machines that are available and you've got this number of jobs that you need to get done, you allocate these jobs to these machines because this is the way that needs to get done. That in and of itself is a highly complex process, but here you're talking about this is a much richer set of problems because, for one, there's just more sensitivity to — There's a wider range of sensitivity to the tasks that need to get done.

In Netflix, there's pretty good failover for a lot of the scenarios, like you can just fail over to the defaults and it's like, "Okay, you've got an index of movies and you just need to be able to play movies," and most of the scheduling that they're doing is like, "How do we forecast what movie to recommend to somebody?"

If you're talking about IoT, you're talking about forecasting, "Okay, is there to be a car in the road in the next 10 feet or is there a safety hazard on my on my construction site?" These are things where you need a much higher degree of sensitivity. What changes in this world when we have these higher demands on the processing systems and the machine learning stuff, because it seems like that we need a different type of scheduling strategy.

[0:08:56.0] OB: I think you're actually aligning a very interesting scenario here. What is key here to solving these kind of issues is exactly what I mentioned which is running the compute on the data at the right place at the right moment. For the scenario of detecting issues or problems and foreseeing problems such as in a car situation or in an industry on situation like a factory floor, you need to have reaction times that are very low. You need to instantly react, but based on what? You have option one which is actually you have sensors locally and you can pre-configure thresholds, and so your device is powerful enough to say, "Hey, when that value goes above that, let's stop everything." It's kind of binary. There's not a lot of refinement in terms of what can be done in terms of reaction.

The other option is to have some bigger unit doing the compute for you based on bunch of data. Then you enter into that idea of, "Hey, let's push all the data from all these vehicles, for example, or all these robots in the factory floor. Let's push them all to the cloud and have some rich machine learning algorithm being trained based on all that data, and then we have a model that can be applied to real-time data and detect these anomalies and eventually react.

However, you still have in that situation the round-trip to the cloud which is a problem in terms of the latency, in terms of reaction times. Now, with technology such as Azure IoT Edge we just announced, you can think of starting distributing that. If you need to have a specialized model on specific data to be executed, to be run as close as possible to that data, you can train the model in the cloud and then deploy at the edge on the devices.

This is actually the type of complex scenarios that we are trying to solve allowing a continuous training of these model based on a lot of data, it goes to the cloud. Then leveraging these compute units that can be run on the devices at the edge to execute these very specialized models that have been trained up there and reduce latency time, have reaction happening at the edge, but having a reaction is no longer based on thresholds and I would say dumb time limits or preconfigured and hard to update type of limits. This is something that actually will be happening on very smart algorithms that will be executed super fast because they're very much optimized.

Then the other thing is how do you manage all of that? Because updating software on a device, we know how to do that, but it's complex, it's hard, it's long, and it's tedious, and you have all the security aspect of it to take into account as well. You don't want some random software to end up into your car controllers, right?

At the end of the day, beyond the capability, the ability to run compute services at the edge, what actually is also a big value and help the customers right now because that's what they need and looking for is a way to manage that intelligence, a way to deploy it, a way to update it. All of that in a secure way and all of that control through an interface which is very practical from anywhere which is the cloud.

Actually, there's two aspects, the wrong time on the devices capable of running that intelligence and all that orchestration and management of hosted devices and which entities are running on this devices that is key. Basically, that is exactly what developers are looking for and our customers are searching and in need right now.

[0:12:39.2] JM: I saw you at Microsoft Build and there were a number of presentations about manufacturing and construction companies at Build. They were talking about IoT, and these

companies are typically outfitting their machinery with lots of sensors and they want to do predictive analytics. They wanted to predictive maintenance. They've got a lot of old technology and they're often integrating newer technology with the old technology. What's the process for deploying something like an Azure IoT solution? How do you onboard them?

[0:13:23.2] OB: We actually are seeing that and we have different names, so that's the area where some customer already have an existing, we call that the Brownfield, and when they are going new with a new type of project, it's Greenfield. Brownfield is most of the situation when you think today about typically the industrial automation scenarios. When someone builds a robot and someone buys a robot actually, they're expecting that to be working for the next 10, 15, 20 years. They don't want to change that every now and then.

When it comes to these kind of scenarios, these devices locally already have tons of sensors. They are basically feeding into equipment such as PLCs that are kind of local computers that do all that work of processing real-time data and taking action based on complex configuration of thresholds and so on.

The first step for these kind of customers is to apprehend the fact that you can integrate into your IoT infrastructure, existing infrastructure, a gateway or a set of devices that will act as they're interfaced to a broader system. When I say broader system, I think about the cloud as one entity to be to be connected to but you can think of as well about private cloud solutions, on premise type clouds or their own infrastructure, because actually there's also the need to eventually hookup these machines and these infrastructure into their own line of business applications to trigger maintenance, optimize their customer relationship management and things like that.

The first step is really to look into how can I inject into that infrastructure smart and connected devices that will, on one hand, get the data from these devices, send them to that cloud entity or back-end, allow for analytics using very advanced services on that data. Then take action by re-injecting commands down to the devices. That interface was Brownfield infrastructure is done through gateways. It's done through protocol translators and things that we have in our portfolio that can help the customers, let's say, cloudify their existing solutions.

Then there's the other type of customers who are actually starting fresh. In that case, it's even easier in terms of implementing a solution because you have to start fresh, and these devices will, themselves, be connected and smart enough to just remember things.

Both scenarios, both type of approaches are addressed here and we definitely have in mind the fact that customers are not always starting fresh. That Brownfield type of scenario is key. It's actually very much more than just adding a Wi-Fi chip on an existing machine. Security is something that is customers have to take into account very seriously, because they were on the close environments, that were like totally safe in their environment and we're able to have full control. Now, they need to think security. They need to lock things down. This is where they can benefit from an infrastructure and tools and SDKs that give them that security out of the box.

I would say this is definitely the two angles to look into how do you extend an existing solution and how you look down securely all the connectivity and so forth.

[0:16:50.6] JM: When I think about manufacturing and construction companies, these are the kinds of companies that I imagine as being slow to the cloud, similar to banking or finance or healthcare where they say, "We're not so sure about this whole cloud thing." What you just said there made that argument a little more rational to me because, essentially, what they would have to do to go on to the cloud is open up some endpoints that expose them to what feels like a riskier situation than just running this, I guess, what they would have run in the past which is kind of like an intranet where they have complete control over everything.

[0:17:36.4] OB: Exactly. Actually, this is something that can be done in a very bad and fast way, which is not securing all of that. We've seen in the recent past that this is not really a good idea. This is definitely one of the factors that makes this industry is not slow but more cautious in terms of adopting the cloud. Another aspect to it —

[0:17:58.6] JM: Because if you get it wrong, you get ransomware.

[0:18:02.4] OB: Well, you get these kind of things. Yes, if you get it wrong, you can get in big trouble. You can lose data. You can actually have men middle type attacks, a thing that actually can compromise your business and privacy and all these aspects of things.

There's that aspect which is kind of slowing them down, but they're getting to it and they're learning that it can be done safely, it can be done securely. For that, they need to have the right tools, the right infrastructures the.

The other aspect to that's, not slowness it would say, but it would qualify that as cautiousness. These customers are very cautious and the other reason is sovereignty of data, where basically lots of them, the core of their business is around IP that they built around data. A good example could be the one we showed at Build, [inaudible 0:18:55.3] which is that company building this high quality and precision metal cutting machines.

What's happening that what they have a business on today, or yesterday, where they had a business on having this very advanced and efficient machines, but they're not the only ones able to do that now. They're shifting their business model to provide services for having a better efficiency at how these machines are used.

Basically, they gather a lot of information from all these machines that are across the world in these factories, on these factory floors, and what they do is they determine what's the best use for these machines and tools and, basically, feed their customers with that information. Suggesting when to use what and where to use that tool versus this other one and how to reuse that tool while it's not using that machine, on the machine down there that is not running at the moment and things like that.

Basically, optimizing workflows and optimizing productions. So the analysis shifting they're model to is still build very advanced machines, but we also sell the service and that makes the value of these machines which is the service that comes with them the last to save —to save a lot of operational money.

They are directly differentiating from their competitors things to these kind of solutions. This is another shift which is very hard for them which is switching their business model and being able to face competition in a different way. This is very hard for lots of industry actually.

[SPONSOR MESSAGE]

[0:20:40.1] JM: Ready to build your own stunning website? Go to wix.com and start for free! With Wix, you can choose from hundreds of beautiful, designer-made templates. Simply drag and drop to customize anything and everything. Add your text, images, videos and more. Wix makes it easy to get your stunning website looking exactly the way you want. Plus, your site is mobile optimized, so you'll look amazing on any device. Whatever you need a website for, Wix has you covered.

So, showcase your talents. Start that dev blog, detailing your latest projects. Grow your network with Wix apps made to work seamlessly with your site. Or, simply explore and share new ideas. You decide. Over one-hundred-million people choose Wix to create their website – what are you waiting for? Make yours happen today. It's easy and free. And when you're ready to upgrade, use the promo code "sedaily" for a special SE Daily listener discount. Terms and conditions apply. For more details, go to www.wWix.com/wix-lp/SEdaily. Create your stunning website today with Wix.com, that's W-I-X-DOT-com.

[INTERVIEW CONTINUED]

[0:22:05.7] JM: For some of these companies, the term machine learning, all that means is maybe gathering analytics, gathering data on, let's say, just some atmospheric check, like the amount of gas of a particular type of gas in an area of a factory and being able to respond to that. Simply, if you have enough of a certain gas in an area, maybe you send a signal to shut a door somewhere or open the vent somewhere. Maybe it's learning to predict what are the factors that lead to that kind of gas being in the air. It could be something very simple like that. Then there are much more complex models that these kinds of companies would want to build. What are these countries asking for in the way of machine learning?

[0:23:05.2] OB: It's actually different type of things. The first one they're asking for is because machine learning is complex. Let's just say it's a hard. It takes data scientists to determine models and work on them. The other aspect of it is of the complexities that data scientists might not know a domain of application of data. There is a need for collaboration between what the industry is used to and the data they have and they own and they know how to use it and what it means and so forth.

The data scientists needs to makes sense with that data through rich algorithms through machine learning runtimes and compute powers. At the end of today, the thing that the customers are asking is simplicity in the complex work that machine learning is about.

First thing, a set of things that they want is a set of simple APIs, and when you think about cognitive services on Azure, this is exactly what it is. As in, if you want to detect anomaly on timestamp data that is sent over, you can actually apply specific anomaly detection algorithms calling one API, one REST API, you send a buffer of data, and this model, this machine learning entity, will learn from the data you're sending and will respond with anomaly detection alerts. A long time, this algorithm will just learn, keep on learning. The more data you send, the more it will learn, and on its own, will be able to determine even better the anomalies.

You have other algorithm that could be accessible through APIs such as face recognition, voice recognition, and others. These are just the samples, the example story that come to mind and talk to everyone, but there's plenty of other ones. Having machine learning service in the form of cognitive service in terms of simple APIs solves a lot of the problems customers have.

The other one is the possibility of running these algorithms somewhere else than just in the cloud. Actually, this is new. This is also something that might have like slowed down people adopting these algorithms. The ability to train these models, train these APIs, in an infrastructure such as a cloud, super easy, you have plenty of nice tools and it can benefit from existing libraries and APIs. Then you extract some of that, the specialized part of it, a thing that you really need for your machine or your solutions and then bring that down to the edge.

This is actually — The other thing that customers are looking for in machine learning which is making their devices smarter, but not just thanks to a cloud entity that does all the work, also by running specialized algorithms down at the edge.

[0:25:56.2] JM: Okay, you mentioned a few really interesting things there. The first was that this was like this video that Microsoft showed during the keynote where when you're referring to these cognitive services, there are a couple of scenarios that Microsoft displayed at build where

it was totally new to me and this is a really interesting idea. Basically, you have cameras throughout your construction site or throughout your factory.

If there is a situation that needs fixing, then the cameras can detect it. For example, “Oops! Somebody left that bandsaw out,” and if that's in the wrong area, somebody is going to trip over it. That's a safety hazard, and the cameras can identify that safety hazard and notify somebody, “Hey, pickup that saw somewhere.” You can also identify people on the construction site who you want to remove from the construction site. That’s one part.

You mentioned the cognitive services thing, and I think this is slightly related to the IoT Edge stuff, but I think we should talk more about the machine learning stuff. Anyway, I just wanted to comment on that, because you mentioned the cognitive services stuff and I had seen all that facial recognition, image recognition stuff that Microsoft showed, basically, how you can apply that to something like a construction site, and I thought this was just fantastic. You see all these stuff like, “Oh, here's how you can apply better filters to Snapchat or to some Facebook thing.” It’s like, “Cool, that's a nice use of facial recognition.” Here, we’re talking about real time way to respond to potentially dangerous incidents on a construction site. I think that's just fantastic.

[0:27:46.8] OB: As a matter of fact, tying back to Edge, the actual demo and scenarios being showcased that you’re describing, was leveraging Azure Stack, which is an Edge solution to run beefy huge services down on premise. Because one thing that is hard to do was video, like that. Imagine like the scenario where you have all these cameras to secure that construction site. You have to have good coverage in terms of video if you want to have visibility on older tools that are around and all the safety hazards that can be can be seen. Once you have that coverage, imagine the amount of data that these cameras are sharing.

Now, you cannot do that if you don't have smarts and intelligence at the edge, because you cannot stream all these cameras on one single GSM connection or whatever it is, or it will cost a lot of money with like super expensive fiber-optic type of connections.

Basically, now, because you're able to bring intelligence at the Edge where the devices and sensors are, you can still leverage the power of machine learning cognitions service in the cloud because it runs locally, because you're able to execute these algorithms there. Then you can do

these very advanced scenarios. You don't rely anymore on a round-trip to the cloud on very expensive bandwidth usage. Now, we can enable the scenarios that previously were not possible, not from the technical perspective, because actually since video recognition exist and open CV libraries, open-source, doing it is not hard.

The problem is doing it on hundreds of cameras at the same time and determining actual interesting and relevant information and take action on time to actually avoid dangerous scenarios, avoid incidents and go beyond the liability problems that could occur in using such a system, then it requires having things running at the Edge. Hence, the importance of having this intelligent cloud supported by an intelligent edge.

[0:29:58.3] JM: From the business side of things, what's so great about this is this is totally in Microsoft's wheelhouse. The idea of, "We're going to sell you some computers and you're going to use those computers on premise to do your work." That's what Microsoft has been doing for a very long time, and here is yet another scenario where it makes sense for Microsoft to say, "Hey, let's sell you something that works really well on premise."

[0:30:25.4] OB: Yeah, exactly. This is this is not new both from us and from others actually. The model definitely works really well, and we developed a lot our work in the cloud in the later years. We are complimenting that work with something we've been really good in experience with, which is all the Edge activity.

Just as a reminder, my previous life before my Microsoft life, I was a developer in the embedded space and some of the best embedded operating systems at the time were Windows CE and Windows XP embedded, things that actually were working super well, being smaller complementized version of Windows and real time for Windows CE and able to power machines and still have all the security, all the management connectivity and ease of developing apps for.

This has evolved to the point where actually we saw a big interest in working more into cloud, but you can see we're reconciling both and we are actually now leveraging all our strengths. Definitely, it makes sense for Microsoft. In general, anyways, it makes sense to have this intelligent quite actually being extended with an intelligent Edge. This is what the customers are

asking for. This is exactly the type of scenarios that they want to implement or try to implement today.

[0:31:51.6] JM: When you're selling the service — Okay, the idea of a company that has cameras all throughout their factory or all throughout their construction site, do they already have these cameras up and running or do you need to sell them an entire camera system that's compliant with Windows. What exactly do you have to do for that process?

[0:32:11.4] OB: For that's very specific service that has been presented in that demo, I don't know details myself. My understanding at that level is that they work with regular camera that have a descent video quality, because actually what's happening, that they don't require to have a special connectivity to the cloud themselves. They don't need to have a special compression algorithms, because you don't need to compress for sending the data. It's all local.

Basically, basically you can do that with very simple cameras because you don't need to have all of that in the camera, because you need to send the data compressed or already pre-filtered. It's happening at the Edge. It's happening down there. You just send a row feed of the camera to the stack instance that has all these services running and you your analytics on that stream of data down there. It's not saying that it doesn't require some beefy set of servers and computers to be installed down there, but you save yourself from the round-trip to the cloud or from huge amount of bandwidths to be used and you save yourself from having to have very advanced cameras that do treatment themselves with that image.

My understanding is that this service actually will be to leverage very non-expensive cameras. I don't know about situations where people already have cameras deployed, but I think this is definitely something that we have in mind.

[0:33:38.3] JM: Let's say I've got cameras all throughout my factory and all these cameras are feeding video data into some computer instance somewhere that's running Azure Stack, I guess. Is that what's happening? What is Azure Stack doing, and what's the software stack? Is it Windows? What's going on there?

[0:33:58.2] JM: Azure Stack is actually — Think of it as Azure, an instance of Azure running on your own machines. Basically, you have a tinier version of what Azure can do and there's a set of services already supported in stack that run today in Azure that it can run on Azure Stack. Not all of them are available, but as we move forward with Azure Stack, you're able to run more of them.

When you think about Edge intelligence, Azure Stack is one of the option for when you have the infrastructure to run a full instance of Azure. Typically, you can run — These are powered by VM's and the fact that it's Windows, Linux doesn't matter, because that's a cloud instance that you're running locally.

If you don't have that kind of infrastructure or don't want to invest in that, and you don't have that need of running full-blown services or all of them or huge quantity of data, now you can think of Azure IoT Edge as a lighter solution that still enables you, allows you to run analytics at the edge on tinier devices. There's kind of complementarity between Azure Stack and IoT edge where IoT edge is the tinier devices type of targets, and Azure Stack offers you the big guns, I would say, on premise.

In the case of the cameras, we're talking Azure Stack, we're talking about full-blown cognitive services running at the Edge on Azure Stack servers in your factory floor in that case.

[0:35:33.5] JM: Can you talk more about the operating system. I guess I don't know much about what Azure would even be running in the cloud to host that information, host the services.

[0:35:45.4] OB: You basically would have a runtime that basically is based on Windows and that allows you to run platform as a service services. Basically, you don't need to care about VM's and all of that. You can run a VM up there if you want to, but the idea is to offer these services as best, and so you have a platform as a set of APIs, as a set of services there. The fact that they run on that or in other is what's inside Azure itself. It's based on Windows, VM's. You don't have to deal with is the scaling, not just to deal with the elasticity. You don't have to care about the load-balancing. That's what the Azure Stack offers you, which is the peace of mind regarding the infrastructure. It's not like you just have a set of servers you have to manage yourself. You have Azure that does that orchestration, load-balancing and so forth.

[0:36:41.5] JM: If I'm getting all this IoT data, let's say I want to store all of it. Let's say I wanted to store all of my video data or all of my centrifuge data or all of the data on the atmospheric information throughout my factory, what kind of database do I use to store all that data?

[0:37:02.0] JM: I'm not aware about exactly which is supported today. We can look online about Azure Stack. There's going to be storage services supported their and the various databases that are supporting, when you look at what Azure has, gives you a good idea what Azure stack will support. It goes all the way from SQL-type databases, to NoSQL ones. There's that new thing that we call CosmosDB, which is NoSQL type of database that you can leverage.

In the cloud itself, there is data lake that actually is able to make sense of all these various types of databases and data sources. There's plenty of options and, basically, this there's no one that would tell you, "Hey, you have to use that or that." It's very much based on your scenario. What do you need? Is it performance? Is it accessibility? Is it —It's definitely is there available for you to use based on your scenario and your choices.

[0:37:59.8] JM: Okay. When we're talking about training and deploying machine learning models, if I'm a big factory, am I doing some of these machine learning workflow on-prem and some like I'm offloading to Azure cloud, or what's the — Yeah, what would do there?

[0:38:28.6] OB: Usually, if you already have, which is like lots of cases out there, you're on data scientist work in your data because that's your business model, or because you're optimizing workflows and thing like that based on that. Very often, basically today, these data centers use technologies such as R and other ones that they are supported in the cloud. What you benefit from if you start doing that in the cloud is the scale.

Basically, you're not limited to what your local hardware can do and can store. Now, you have the unlimited storage, unlimited compute of the cloud, which means that you can train your models even better because you can augment the amount of data that these models used for training. You can also accelerate at what pace these models are trained.

Basically, you can leverage to cloud to make that process even better, refined these models. You can also leverage other models coming from the libraries that we have up there for cognitive source and others and extend them with our tools. Then when it comes to the execution of a model or the use of a model, because —Machine learning, you have these two aspects of your model, which is the training where you reserve a portion of the data it's coming in for training, and then you have the execution of that model on the data to do that real-time analytics, with predictive maintenance, or these kind of things.

Basically, the ability to leverage an edge entity to run this model that has been trained in the cloud is actually very practical for reducing the latency, the response times for optimizing on the machines. Basically, you can leverage now the two worlds. You can have this world of like unlimited compute and storage of the cloud and then make the most of that very optimized models all the way down to the edge to optimize on something else.

[0:40:32.2] JM: You could also, if you have a network of factories, you could have all the different IoT data from your different factories being crossed and measured against each other in the cloud.

[0:40:46.1] OB: Exactly. You multiply basically — Potentially, you multiply the sources of information. The more information a machine learning model has for training, the better trained it is, because you multiply the situation, you multiply the patterns the models are exposed to. Basically, you multiply the chances for that model to learn more about the data.

Having data coming from lots of places is ideal. He can imagine that he have like patterns that's been identified in one factory that's a way to predict it can have the same problem in another one. You could have a machine that has been positioned in a wrong place in the factory floor because that's next to a big fan that is actually blowing very hot hair on to it and it's going to fail. Then you can determine that, it's not because this machine was having a problem, it's because it was not well-positioned in that factory because the same machine in the other factory doesn't have the problem. The only difference is location in the factory. It can actually, because you gather data from many sources, you can make more out of the data.

The other thing, which is very important as well for the customers is to differentiate their business to grow their business. Doing things such as just selling the metal cutting machine we're talking about to go into a business where you can actually have services sold to your customers with the machines helps you differentiate your business, augment and grow your business.

Because you have that ability of gathering all that information from the various factory floors, from your customers, and then mix-and-match that data and re-offered that mashed data or analyzed data as a service back to your customers, now you're able to offer that value on top of your original business.

[SPONSOR MESSAGE]

[0:42:52.1] JM: Artificial intelligence is dramatically evolving the way that our world works, and to make AI easier and faster, we need new kinds of hardware and software, which is why Intel acquired Nervana Systems and its platform for deep learning.

Intel Nervana is hiring engineers to help develop a full stack for AI from chip design to software frameworks. Go to softwareengineeringdaily.com/intel to apply for an opening on the team. To learn more about the company, check out the interviews that I've conducted with its engineers. Those are also available at softwareengineeringdaily.com/intel. Come build the future with Intel Nervana. Go to softwareengineeringdaily.com/intel to apply now.

[INTERVIEW CONTINUED]

[0:43:43.4] JM: Let's say I'm a factory and I'm collecting all these data on my centrifuges, maybe I can resell that data to some data broker who aggregates information about centrifuges.

[0:44:01.2] OB: That could totally be a use-case or a scenario. Yes.

[0:44:04.8] JM: Pretty interesting.

[0:44:05.7] OB: Actually, the one I was thinking about and describing is actually a scenario that our customers are implementing, which is more about; when I'm selling these machines rather than just selling the machines, let me gather the data they're producing. For example, we have customers that are building coffee machines, like professional brewing machines that are visible in some coffee shops here and there.

These machine, just being locally logging some information is not good enough, because very often they fail for someone to be sent for maintenance or sometimes they send someone to do maintenance on a machine that doesn't need to be maintained. Basically, having these machines starting, sending their data, they're not only sell the machine, but then say, "Hey, you know what? For that's monthly subscription, I'm going to optimize your maintenance and it's no longer going to be about one guy coming every month to you. It's going to be when it needs to come."

The other thing is on the usage of the of other customers, they'll be able to provide some predictive information about your own machine hear and I'll be able to tell you, "You know what? You should turn your mission off at night because actually all the customers who are doing that is just saving that amount of money."

Basically, now you entry to that relationship with your customer, which is no longer just about selling and maintaining machine becomes a joint venture of optimizing the business of the coffee shop and saving money in other areas. That could be a trade and a customer might be willing to go with your coffee machine because it comes with that additional service with it. You have this kind of new type of businesses and relationships that can be built.

[0:45:52.1] JM: The world that we've been discussing with the construction sites or the factories, this is a world where the model of computation is still pretty well-defined. It's kind of a client/server model where you've got the client is the entire intranet of the factory. It's the cameras that are feeding into the Azure Stack server and you've got some sensors around your factory and stuff like that. Then for certain use-cases, the Azure stack on-prem thing is to communicate with the cloud and do some stuff. That's pretty well-defined and I could see like a lot of really good use-cases there. I can also see a lot of room for growth. What do you think of the self-driving car and drone space, because this is a little less well-defined, because we're

going to have these things that are just flying around or driving around where they're going to have intermittent connections, and we've also got our smartphones, of course. This is a different model of edge computing because there is such a dynamic mesh of different devices that are moving around. Are you starting to think about this space as well?

[0:47:09.1] OB: Definitely. I think that the elements that you mentioned are spot-on, which is the new problematics that these areas or are bringing; the non-continuous connectivity, the interaction. You didn't mentioned down, which is pretty huge, is the interaction between these entities and devices, because when you're off-line and typically think about the drones, two drones are flying in the air, they're going to enter to collision mode. While the need to negotiate in terms of, "Hey, are you turning right or am I turning right?"

There is a few things that will need to happen there, basically — or happening, because we are involved, other companies are involved, and we wanted to find a common ground for implementing the right things.

The edge intelligence helps a lot. Think about when you are able to download or to actually deploy on to drones new set of models for that kind of interaction with other entities around them, or if you've learned a new pattern based on the temperature and pressure and wind and location of obstacles, and then you say, "Well, if I could actually have that deployed now, it would prevent like lots of drone wreckage." Then you're able to do that, thanks to the edge intelligence deployed from the cloud.

You don't need the device to be perfectly connected because now it can run that model locally. Then you have the communication. We have more and sensors, more and more technologies that we're looking into. You have technologies about low battery, low consumption networking. You have technologies about very fast throughput for data, which was LT5 and things like that.

Basically, all these new technologies that grew to making these scenarios possible and having drones, having self-driving cars, in terms of pure technology, this is not new. In terms of the means for making it happen, this is where things are new, and the aggregation of all these technologies we're looking very closely. We are working on that's on our own and with partners and with customers to make these happen.

There's nothing to announce here, but there will be no surprise that we are working with these partners in these areas. When you see that we showcase cars, we showcase things like that, it's because we're involved in that world. It's not new. You heard about Ford SYNC in the past, and we'll have plenty of interactions with cars. Now we're working more and more on providing back-end in cloud with intelligence, the predictive maintenance and so forth.

Obviously, we'll come to working on these devices that communicate with each other and are able to interact and create these new types of scenarios that will be crazy to think about today, but it will be totally normal for our kids in a few years from now.

[0:50:12.7] JM: Help me expand my understanding of the kinds of businesses that are being built with IoT related technologies. Can you talk about some of the success stories or some of the interesting cases that you've worked with on the Azure IoT team?

[0:50:29.9] OB: There are some very interesting ones. I remember like one around working with Schneider on optimizing the energy consumption on remote sites in Africa for schools and other buildings like that.

In situation, there are complex and there are actually remote and hard in terms [inaudible 0:50:57.4] would say, in terms of environment. Having the ability to analyze the data that is generated by — In that case, was HVAC systems and all the sensors in these buildings, you're able to optimize the use and the consumption of that rare energy down there and allow these buildings to be used by inhabitants for storing, food schooling, for all these kinds of things. You have this kind of like very nice looking type of project.

Then I can think about scenarios that are more practical, I would say, such as the [inaudible 0:51:37.6] example where a company that was "just building high precision industrial machines," now they're selling services along with them. They expanded their business and they went on to that, thanks to the IoT scenarios that are now possible with connectivity, with the smart intelligent clouds and so forth.

We can list a lot of them and they're very often in the vein of, "Hey, I had this bunch of devices. They were pretty dumb, doing nothing." Now, because I'm able to make sense of the data they're producing, now I can do A, B, C and basically I'm able to do new things, thanks to that.

I was always making fun about, "Hey, IoT is that new thing." It's not new. We've been doing connected devices since 20 years ago. We've been doing smart devices 10 years ago. We've been doing connected devices 5 to 10 years ago. The difference right now is that we find a name for it and we do have machine learning AI that are coming into the fray of these scenarios. They're bringing that intelligence that allows making sense out of that and taking action which is that second step. That brings things back to the human, that interaction that we need these machines for these scenarios to be viable.

This is very much the way I see that the future going towards, which is having a richer interface with the humans and having actually the humans taking greater part in these scenarios than we think, because we very much often think about IoT is going to replace humans, going to kill jobs, and I'm convinced on the other hand that is going to be the opposite, because we actually will optimize that interaction. We'll make the human more efficient, and we'll make that interaction more beneficial at many levels.

This is my vision about the future. I'm an optimistic, and I love what I'm seeing. I think it goes superfast. It's very impressive how technology in that area is growing and how we are seeing now the merge or the overlap of these technologies. AI on one side, IoT developing on the other one, he communication technologies on the other one. All of that comes into place.

There's an area we didn't mention which is actually totally integrated, which is all the mixed reality and augmented reality. There are domains that totally align complement go along with IoT and AI. If you combine all of that, you have what our very near future will look like.

[0:54:26.5] JM: Where does that take you?

[0:54:28.2] OB: Me personally or —I think the movies we're seeing are not even close to what is going to be. The seamless interaction with your environment and with things that are not even

like close to you, whether they're devices, whether they're people, whether they are services, all that interaction will become more and more seamless.

Where, honestly, I see things going is to a place where these put together will make our lives easier and simpler and will allow us to spend quality time on the important things. I like to think that where we're going is — It might sound a bit cheesy, but towards a better world based on the fact that these systems will support the humans to do better with everything we're doing.

[0:55:26.1] JM: I'm with you. I agree. I think that's where we're going to. Olivier, thanks for coming on Software Engineering Daily, it's been a pleasure talking to you. I enjoyed meeting you at Microsoft Build.

[0:55:36.1] OB: Thanks for having me and I hope we talk soon.

[END OF INTERVIEW]

[0:55:39.1] JM: Gatsby is an API for running promotions. It's the slickest and smartest promotion platform. If you want to run a promotion for a discount or some sort of call-to-action campaign, Gatsby is the choice for you. It enables customers to increase conversion through elegant feature-rich promotions that can capture email addresses and social profile data early in the conversion funnel. It increases your brand's social influence by helping your customers follow and share your brand while identifying the customers with social reach for follow on campaigns.

Gatsby is opening a private beta of their API and it's useful for developing an in-house promotion or connecting the user data of your customers with the Gatsby app. You can do A-B testing on multiple campaigns. You can automate retargeting. You can generate promo codes and geo-fence promotions. There's lots of other features, and if you're interested in trying out the Gatsby API for your promotions, contact api@thinkgatsby.com. That's api@thinkgatsby.com. Thanks to Gatsby.

[END]