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 using Azure to transform their business. You can find our team online at aka.ms/indxp or on Twitter at Industry XP.

[Jess Panney 00:00:33] has over twenty year's experience helping companies succeed through the smart use of technology. He's spent most of his career working for leading Microsoft partners across the UK and Australia, and is now principal at [Engine 00:00:52]. At Engine, he works with clients to help them transform into modern, data-driven, Cloud-first organizations.

Jess, welcome to the show.

Jess: Thank you very much. Pleasure to be here.

David: You shared some stories with me that I'd love to have our listeners hear too and we are going to talk about data science.

Jess: Yeah, that's right. It's an area we've seen in a massive explosion over the last couple of years, especially. Organizations obviously collecting vast amounts of data. They've now got the capability to do so very cost-effectively with platforms like Azure. People are now coming to us saying, "How can we use this data to answer some business questions that we weren't able to do, certainly, in the time frames that we could before."

So, yeah, there's a lot of excitement in this area at the moment and I think Azure is really sort of helping us to exploit that potential, really.

We ran an experiment recently. So we work with organizations to work with their information, try and answer these interesting business questions. There was one particular one that I found particularly interesting. There was a real estate problem in the UK, we'd gotten a new sort of disrupter on the market where people are coming onto the market to offer services for selling their homes online or their business online, or all sorts of other services and people's possessions in new and innovative ways.

When you're talking about real estate, we are talking about large, high-value product. Organizations typically won't just transact through an online-only basis. They will require some agent or some person to go round and assess the properties, assess the item that's being sold, in order to value it, in order to describe it, take photos of it, potentially. And the problem with that is that it's a very expensive process, getting people around the country, shipping them over to people's properties or to various locations to do this. That's an expensive operation. When you're offering to sell those products to other customers online, there's a lot of tire kickers, there's a lot of people coming around saying, "I'm just interested to know the value of my item or my property" or whatever that may be. And so that's taking up a lot of time.

So, really the business problem challenge was, we want to be able to know or preempt which customers will go on and transact and list their item online, or before this organization would invest in sending agents through to their homes. One of the ideas was actually that you could change the sales final effectively, based on the likelihood or the customer propensity to go on and proceed with that particular transaction. So it's really exciting, it had potentially massive implications for that particular organization in terms of how it would operate from a profitability point of view.

So, we looked at Azure for this experiment. We thought, okay, what information, what open data, what open source data sets do we have? What data sets do they have internally that we could use to leverage this? And we went to them, we said, look, what's your hunch, what's your hypothesis that you wanna prove/disprove? And they said, hey, we've got this data internally that we know about past customers, but we've also got these open source data sets which we use, which if we pull in, we think might that actually unlock the secrets, that actually would really help improve the model.

So, we looked at Azure. We imported their data onto the platform. We used Azure Machine Learning in this particular case to really quickly run a very quick, fast experiment. It was actually two experiments which we ran over a period of two weeks, so it was one week iterations. We used Azure Machine Learning Studio to wrangle the data into a format that we could use for the experiments. We then ran the experiments using a number of different models. And very, very quickly, we were able to establish whether their success criteria was going to be met or not. Whether the hypothesis was going to be met or not.

And what was really interesting is that the data set, this open source data set, they felt had all the secrets to their particular problem. We found that when we ran the model on that data set, actually it came up completely 50/50, flip of a coin sort of model. We ran the experiment again using their internal data set. We had some success on that, but just not enough to be able to progress with that investment. We didn't get a strong signal from that. But it was really interesting that the initial kind of, we think this is gonna work for us, actually turned out not to be the case at all. And the Machine Learning Studio and Azure itself really helped us get to that point a lot faster than otherwise we would have done.

David: How is that so? What parts of Azure did you take advantage of to run these [crosstalk 00:05:46]

Jess: So a big part of it was basically getting the data off of the platform. They had a whole bunch of information, let's say internal information from their internal systems, which we had to pull and we had to ingest that quickly. So we pull that in, into Azure Blob Storage. We use Data Factory to actually get that in. We also used these open sourced data sets and again, we used Blob Storage as basically the mechanism to store that information.

We then used Azure Machine Learning to then, as I say, to wrangle that data into the format we required for the experiments. And then we executed it obviously within Azure machine learning. And the great thing about Machine Learning Studio is that, not

only does it give you a nice easy path for experimentation, it also gives you a really fast path to operationalizing the solution. So, had we found a model that worked really, really well, and we've done that, obviously we've worked on other projects which we have actually operationalized. But this particular one, had we found it, it's very easy to turn that model into an end point that we can call from our pipelines, our other processes, from potentially, in this situation, the company's website. When they're actually registering the customer, they could have ran this model and said, "Alright, what's your propensity to proceed with this sale or proceed with this transaction?" And then we could have made some really cool decisions off the back of that in terms of where we take them next.

So it's a really great platform for really quickly not just doing the experimentation but also doing the operationalization side of things.

David: If I understood you correctly, Azure Machine Learning actually created a web service for you to call to push training data into your model and to have it evaluate data?

Jess: Exactly, yeah, exactly. And to give us the feedback we needed to make sensible decisions based on real data, based on actual evidence rather than, I suppose the approach that a lot of organizations, a lot of people make, which is, "Let's just chuck some data into a model, let's see if we can come up with something." And then they look at the results and then they look at it and go, "Yeah, well, we might be able to use that."

So what we try and do and what we're really, really clear on with the customers that we work with is that we have to put the scientific method around this. We have to be very, very strict in terms of what are the success criteria, what's your hypothesis, what are the success criteria, before we embark on the experimentation. And that helps to eliminate a lot of the cognitive biases, a lot of the people just trying to make the results fit their hypothesis after the result. And really what that allows us to do is allows us to work in short, sharp iterations, if it were, and to then change direction or make sure that we are always working on the most valuable thing for that customer.

David: That makes a lot of sense. And you're gonna have to make several passes at processing data before you're confident in your algorithm and such, right?

Jess: Absolutely, it is an iterative agile-based process. It's taking the scientific method, which is what data science is all about. As developers become more familiar with this type of approach to solving problems and starting to use Machine Learning, they are coming from a background that is more agile, that should be an evidence-based process. And really, what we're doing is we're just pulling those two worlds together really and saying, "Let's take data science, let's make it accessible to people. Let's make it accessible to developers. Let's make it accessible to people who wouldn't normally look at solving these types of problems because they're just historically quite complicated."

And really, these platforms, like Azure Machine Learning Studio, or some of the other, maybe higher-level services in Azure like some of the cognitive services, what they really

do is they just take a lot of the pain away and all the smarts is now kind of taken care of for you really. And you can just, with some level of knowledge obviously, a lot of this you can't come in completely new, but with some knowledge, you can actually run some very successful experimentations very, very quickly with a lot lower costs than traditionally you'd want to do, normally.

David: I'm gonna go back to your real estate example. Were you able to take an algorithm off the shelf or did you guys have to develop that for yourselves?

Jess: No, we used the algorithms ... actually, Machine Learning Studio comes with, it's kind of like a bit of toolkit, really, in terms of building blocks that allow you to run experiments. And those include standard models that you can effectively just plug your data in. You can tune them, actually. What's really good about Machine Learning Studio is it's at that level where you don't need to know the inner workings necessarily, but you can configure the parameters to get the model to perform in the way that you'd really need it to.

And in fact, in some cases, you can get it to also tune that model for you as well. It has the capabilities that do actually allow you brute force a little bit in terms of running various scenarios through the models, trying to find the best parameters so that you could find a high-scoring model.

So yeah, it really does help us work really, really quickly in that regard because we are working with complex models here. But a lot of it's hidden away from us and we can just focus really on the experiment, and not the mass. And I think that's what's really powerful. And also what we can do and what we've found is, it's a great way of communicating to people who are not necessarily data scientists the method we're taking and the meaning of the results.

Azure Machine Learning Studio is a very visual tool, so it provides you with a canvas that you can almost talk users through or people through, your stakeholders through, and say, "Okay, this is what we've done and these are the outputs." And people kind of get it, they kind of understand, without necessarily being experts, but they can understand with a little bit of explanation in terms of what the results mean, and what we can do next, and you can get some really good conversations going in terms of, okay, well, these are the results, where do we want to go next with those results? Do we want to pivot because we're not getting what we want? Do we want to try something else? Or do we want to carry on or maybe try and enrich that with some more data? Or do we just go ahead and operationalize that quickly and get some immediate value from that?

David: You've talked about that as data science as an agile method.

Jess: I think so yeah, absolutely. I think it's one and the same thing. It's evidence-based decision-making with fast feedback loops and continuous feedback loops. And that goes beyond just the experimentation phase. That goes into the operationalization phase, which is, you know, how do we go and actually turn this into something which runs at

the scale and the performance we require? And actually, that's very much as part of the experimentation as anything else.

We've had situations where we've built successful models in Azure Machine Learning Studio, but we couldn't get the performance requirements through. So we operationalized via the built-in web services or implements AML provides us to call to actually score data. We found it wasn't actually providing the performance [inaudible 00:12:56] that we required, and that's simply because it's a managed service. There's trade-off in terms of the amount of sources you have.

What we did is we actually ended up taking that model and it was actually, mostly written in [R 00:13:08] actually. And we ended up hosting that in Data Lake Analytics. We got much better through [inaudible 00:13:15] running that. So, Data Lake Analytics is part of the analytic suite which allows you to run very large workloads over massive volumes of data. But it actually has this really good option to bring run time, so we can actually run our models within the nodes, with our workloads. And that gives us massive scale opportunities.

That was a great example of using AML to kind of help with that experimentation phase and take switching over to another technology based on evidence, based on the fact that we were getting the through-put we required, moving onto another option within the actual ecosystem. And there are lots of them. Now we're really compounding a reason why we are such fans, I suppose, of Azure is that it gives us options. But again, that was delivered in a very agile kind of approach in terms of, try something, try the cheapest, easiest thing to do to prove/disprove whether it works or not and then move on to the next one.

I think agile is really interesting cause it flows through the entire process. And then of course, you've got the other end of the spectrum which is, once you've got a model, once you've trained it, you've gotta keep retraining it, right? You've gotta keep testing it against the data you have, and that is just another feedback loop. That's just another agile process that you have to implement to make sure that your model is still scoring the way you require it to for the problem at hand.

So, it's kind of an interesting theme, I suppose, for everything that we're doing in this space.

David: Do you go back to your customers for new data to train your models?

Jess: Yeah, well really it should be part of the process. So the output of the model and the behavior that, so in this particular case, we would change the path a customer was given through a solution. Really, the output of that path, or the actual result, what actually really happens over time, we can feed back into the model. And that becomes a continuous learning loop. A constant feedback loop to see what's happening, because actually by introducing a change to the system, we're changing behavior and actually that might have an implication for the model because it's cause and effect. If you change

something, there's a consequence to that, and that might change how we proceed going forward.

So, we always have to bear that in mind. You can't just train a model, call it done, put your hands up, say, "That's it, we're good." Because things will always change. Environmental factors, your solution, there's a whole change. Customers' opinions and behaviors will change. We need to make sure we're always considering those. And that's a big part of the whole data science process really.

David: Yeah, I suppose so. And the prime indicators that you use in your algorithms may also change.

Jess: Exactly, yeah, exactly. It's a process which we find is not necessarily obvious for people. It's a new way of thinking about solving problems, and as I say, as people who aren't traditionally from a data science background, can become more and more involved. It's really important that we consider the data science process as almost like the most important thing of this. The technology allows us to get there quickly, it allows us to execute on what we want, what we're trying to do. But, ultimately it's all around process. It's all around getting people to define what they want up front, getting them to run short, sharp experiments, making decisions off the back of evidence, and then continuously repeating that based on the data that you have available to you, really. Wherever that may be.

David: And a lot of people think of what you just described as being the essence of agile methods. And it really boils down, if I understand you correctly, to the scientific method being applied to this process.

Jess: Exactly that, yeah. And it sounds a bit too obvious. It's got the word science in the name, right? People talk about data science, they talk about Machine Learning. It all sounds a little bit obvious but, actually it's one of the things that I think a lot of organizations struggle with, especially to transition to the Cloud. Cause this is new stuff that people aren't traditionally used to doing. They never had the ability to operate at this scale. They never had these tools available at their hand. They'd never had the amount of data that they have now, and it's a bit of learning curve for people and it's a bit of a journey to get the discipline that's required to be able to run these projects.

And that's what we're finding talking to people, is that people love technology and everything else, but they're so keen to try and get something out the door and say, "Hey, this is our first ML project," or whatever. They often forget about the fundamentals and it's the same with software development. It's the same with agile software development. People, they forget the basics of just being true to the method, really.

David: And that makes all kinds of sense, because it follows right in line with the story of agile that you're telling.

I understand too that you were involved in an anomaly detection?

Jess: Yeah, so that was an interesting one as well. I touched on it maybe slightly earlier, but I think probably worth just touching on it a bit further cause that was an interesting one.

That was about taking streams of data around people's viewing habits. So, basically taking television viewing habits in terms of what people are watching, how they're watching it, information about that whole experience they have around that, and looking for anomalies in terms of the telemetry that we receive from that. It's a fascinating scenario, because a lot of the time, you now what you're kind of looking for. A lot of the time you're kind of trying to build models to try and find what you know you're looking for. Other times, it's like just, we don't really know. Something can go wrong. Something might look a bit odd. When things start looking a bit odd, we need to know about it so we can start [recausing 00:19:11] it.

So it was a very different type of model we developed. It was a lot more high through-put in terms of the amount of data we were processing on a daily, weekly basis. And this one, we ran the experiment again. We actually ran this on ML Studio as well. This was one we actually ended up operationalizing fully, but then actually ultimately moving it over onto [inaudible 00:19:35] eventually.

It was a fully end-to-end solution. We actually took an open source R model for this one. You can go and you can write, use the built-in modules, models available in their studio, but sometimes they don't quite fit the mark. Sometimes they don't quite work the way you want them to, or maybe they just don't perform very well.

David: And you can make your own R script and run them right inside [crosstalk 00:20:05]

Jess: That's exactly what we did, although we didn't even have to do that for this particular case. We took a well-known open source anomaly detection model, we hosted it in ML Studio, we ran the same process. That's really important, just because you're pulling on off an open source R model. Just because you're using that doesn't necessarily mean it's gonna give you the results. We ran the same process, we ran it, we checked it, we tuned it, we validated it against our acceptable criteria. Can we find the anomalies that we expected within a training set?

And then we were able to operationalize that. We ended up using the [inaudible 00:20:40] points to operationalize that. We used Azure Data Factory to operationalize it, so we were pulling in data from an external third party on a daily basis. We were using HDInsight to aggregate that data in a number of different ways. We were then sending that information, scoring that information after it was aggregated into Machine Learning model, and then we were getting the results.

And as I say, what we found was, that was great. That was working really, really well. We ended up moving that away from ML Studio onto Data Lake Analytics. We were able to do that really quickly and really easily. Effectively it was just an R script. We lifted up and hosted somewhere else.

So, that was a really interesting one. We powered off the back of that. We were using [Power BI 00:21:23] really to highlight those anomalies and to get the insight to the people who were interested. So, it was a full on end-to-end solution which, yeah, was very successful. It reduced certainly the overhead of the organization who was very used to spending literally weeks plowing through Excel spreadsheets to try and find the insight they required into something which was just there, presented to them on a daily basis in terms of, this is the area you want to look at. Something strange is happening.

And then of course, they're only showing you the things that you're interested in rather than the continuous noise that is, this is just normal. Rather than showing them just normal, we should them the peaks, the troughs, the things that were anomalous. And that really, really helped them and really helped speed up their time to resolve issues on the platform.

David: And that observation by your client took place in Power BI?

Jess: So, yeah, that took place in Power BI. So yeah, we've done a fair amount in Power BI. We find it's a really powerful visualization tool. Very powerful in terms of trying to get people to move away from the Excel spreadsheet world, or the more traditional report, and just showing the insights and allowing users to interact with that and to visualize that in really interesting and different ways. So we've had a lot of success over the years with Power BI and yeah, I'm a big fan.

David: You've mentioned Data Lake four times.

Jess: Have I?

David: You have.

Jess: It's, yeah, so ... Data Lake Store and Data Lake Analytics are great products. We're firm believers in using the simplest options, or using the options where we don't have to think about the details. I don't really want to spend my day, I will if I have to, I don't really want to spend my day looking at servers or making sure that VMs are running properly and patched and everything else. I like to let Microsoft do that for me.

We have the option, obviously. We can do that if we want to and that's always our option for some problems where you can't get away from that. But what I love about products like Data Lake, Data Lake Store, Data Lake Analytics, and also products like AML is it really just takes that away from you. You focus on the problem you're trying to solve and let's say, let Microsoft take care of all the heavy lifting of the infrastructure.

David: So what does Data Lake provide to you? Because in your first example, with the real estate organization, you moved to Data Lake and it sounded like you did that for right time reasons?

Jess: Yeah, actually, it was the anomaly-detection one we did. We did exactly, yeah, you're right. The real estate one, we actually pivoted. We chose not to continue with that

experiment completely. Well, we ran the experiments, we chose not to continue down the path that they wanted to just simply because the evidence wasn't there and our ... we were very matter-of-fact about it. We said that, you know, clearly there's no evidence here to proceed on the basis you're not gonna get the return you wanted. Let's look at something else.

So that particular problem was operationalized in the end. The anomaly detection one absolutely was, and that was, you're right, it was a right size thing. It's about exploiting the services to be able to get the performance, get the scale, get the behavior at the right cost, you know? And so that was very much a case of, let's look at the different options we have available to us. And, as I said, Data Lake Analytics is incredibly flexible, and for the type of problem we were solving, in terms of the type of model, it was very easy to partition the data up. Because Data Lake Store really is about storing vast amounts of information that is the potential source for intelligence and all sorts of interesting modeling opportunities.

Data Lake Analytics is the brains, I suppose, behind that. So, it's the computer that allows you to run those models. So not only have we kind of got the storage capacity of being able to store vast amounts of data and being able to aggregate it and query it in ... vast amounts of unstructured data. We've got the tools to then run powerful workloads. And as I say, it's flexible enough that we can run our own R in there. We can run Python, we can run [C-Sharp 00:25:57]. It's a very flexible, very, very powerful piece of [inaudible 00:26:03] data problems, really.

And it really does feature in a lot of the projects we've been working on recently because, I think I said at the beginning, these are big data problems. People are having these genuinely large data problems these days and it just allows us to react quickly and produce results faster than we otherwise would.

David: Well, Jess, unfortunately, we've run out of time. You're a great storyteller, I'd love to hear more but we just don't have the time for it.

Jess: It's been a pleasure. It's been good fun.

David: It's been an absolute pleasure for me as well. Thank you so much for joining us on the show.

Jess: No problem at all. Thanks for having me.