Podcast: Scaling the AI Advantage by managing unstructured data

Andy Packham
Hello and good day. I'm Andy Packham, Chief Architect at the Microsoft ecosystem here at HCLTech. And here we are in another episode of Elevate, our podcast around the intersection of innovation driving real business value. And once again, I'm super excited. We've got Srini and Jeeva. They've agreed to come back again. And we've been talking about data and how data certainly is driving this AI revolution.

But one thing we've really not covered yet is we live in a complex world, incredibly messy world. And actually, most of our data is super messy. It's full of value, but it's really unstructured. And what I really wanted to do today was talk about what is this difference between structured and unstructured data? What's content? And is content the stuff that lives in our email, our Excel sheets? video streams, you know, even this podcast, this content. And how do we manage all of that and think about that? It adds uniquely more complex problems, but I think there's uniquely more value in getting that.

So, you know, both Srini and Jeeva, you've introduced yourselves in the past. So I thought we'd just like dive straight into this subject and kicking off and thinking, you know, Jeeva, you know, what is, is there a definition for unstructured data?

Jeeva AKR
Absolutely Andy first of all great to be part of this ongoing series and I enjoy doing this with the HCLTech team both Srini and you so number one I think I want to provide the context uh just a refresher of where we started and how we went along with this entire webinar one as I mentioned in the prior sessions if you think about it I told that In all the 30 years, most of our customers, any time when they did analytics, it was always focused around the data that came from applications which are centrally hosted applications like ERP, CRM, SEM, and all that. Basically, we call them as structured data, right?

And but in the in the in the last five years time what we are increasingly seeing is that the delimitation of both the analytics and the boundaries of the data is actually shifting and the reason for that is this number one one. There are the boundaries of final analytics is actually shifting to from operational reporting focused on yesterday last week last month last year. to more real-time analytics based on the data that comes into right now to do the predictive analytics. And then the second thing is about AI ML applications consuming data.

So that's a big shift that is actually happening as we see it. And then the second thing is that it is also the data. The data is not getting limited only to the structured data. And the primary reason for that is that one we all have seen the statistics right so eight ninety percent of the entire wolf's data has been created in the last four years time and eighty percent of it is actually coming from unstructured data and you already called it out right so the unstructured data is not that it's uh it's easier to whites uh you know white space it rather than trying to provide all the definitions but any data that doesn't conform to the rows and columns that we see in the structured data.

That means it is audio files, video files, and any IoT data that we collect from all the devices, interpersonal communications that we have between ourselves in a common forum, the security logs, the log data, telemetry data, all of it combined together. If you

think about it, in an unstructured data world, we are estimating that telemetry data, which is actually one of the significant part of the unstructured data, which is coming from the data that is produced by all the devices is going to be one of the fastest growing data segment of all data types, right?

So when we talk about the importance of unstructured data, I think it's very critical for everyone to understand that more and more of our customers are actually looking to see value in trying to combine both the structured data and also unstructured data when they are actually looking to do analytics or running, trying to do AI ML applications to train data.

So if you think about the AI models that we are seeing today, In the world of Airbnb or Uber, everything is going to be based on machine data, right? So you can't leave the unstructured data anytime when you are trying to do the AI ML or real-time analytics. So increasingly unstructured data is going to govern the space and it is going to become part of the enterprise landscape as we speak. And I think you're right. Customers need this. There's all this data they've collected. Getting that value out from that is critical.

Andy Packham
So, Srini, you're in so many conversations with customers here. When it comes to unstructured data, what are the sort of things we should be watching out for, being careful about?

Srini Kompella
So thanks, Andy. Again, I echo what Jeeva said. I mean, this series of podcasts is a real pleasure to talk through some of the most critical challenges our customers are facing in terms of A, managing data, B, monetizing data, and C, scaling AI. So when it comes to unstructured data, as Jeeva said, majority of the data that exists in the world today is actually unstructured.

So when we think about managing unstructured data, arguably you could convert unstructured data into structured data because that's what people have been comfortable with in terms of producing analytics or AI over the last couple of decades. But that's A, going to be significantly cost prohibitive and B, it's inefficient.

So today we've got technologies where you can mine insights out of unstructured data, whether it is audio files, video, You know, documents, emails, text, sensor data, machine data. You know, you've got technologies today. You can actually uncover insights in near real time. You can, in fact, even apply AI in near real time.
Think facial recognition as a simple use case.

So when it comes to managing unstructured data, you know, we've got to unlearn the things that we have created. in terms of managing structured data. So there's no need to think about getting all the unstructured data into a single repository and then hope and pray that there could be some meaningful insights coming out. So that's point number one. It requires a very different thinking.

The second aspect is that because the volumes of unstructured data is so high, there needs to be a deliberate strategy in terms of how to monetize the unstructured data. Otherwise, it's cost. And it could be value in the future. So what that means is that the data strategy then needs to delineate between what is valuable data today, which then needs

to be managed differently compared to data that unstructured data that you want to store for potential future use.

So the ones that you want to store, you want to optimize your storage, obviously, reduce your cost. the ones that are going to create value, you got to start with what is the business value and then work backwards. Let's take some examples. Let's say we're talking about video analytics. Think about a retail store. There's something that we've done with a client where in the retail stores, you have cameras that are set up obviously for security reasons. but that video information can also be used to determine optimal workflows.

When I say workflows, the human flows, to determine shelf optimization, shelf spacing, to determine the checkout counters, all in the interest of improving the consumer experience in a typical retail store. Now, this processing can happen near real time as well. And it also be applied holistically across retail stores for a particular chain to improve the overall consumer experience plus build consistency. Now, in order to manage this unstructured data, you don't necessarily need to move all of the video data into a central location, et cetera.

You can gather intelligence, quote unquote, at the edge, which is at a retail store. and then build the aggregation across the retail stores to come up with standard patterns that apply to most scenarios. So that's one example in terms of how you start with an objective of improving the consumer experience and work backwards to figure out how you're going to manage the data and hence set up the data architecture. Another example, you know, water supply. I mean, you know, water is becoming a precious commodity, fresh water, drinking water.

So you want to minimize the wastage. You want to also minimize, you know, bursting of pipes, et cetera, due to increased pressure, so on and so forth.
So we're actually working with a water supply company in Australia where, you know, Jeeva talked about, you know, machine streaming data, right? We're actually using the streaming data. Now combining it with social media and call center data to triangulate zones that are high risk for water leakages and water loss.

This is actually helping this company in Australia minimize water loss exponentially. And at the same time, improve the consumer experience because you no longer have those significant variations in the pressures of the water supply. Plus, you don't have those water leakage problems.

So there's again, now we're getting to the next level in terms of how to manage unstructured data, which is, you know, cross tab across data sets. So here we're talking about taking the sensor data, cross tabbing with call center data, with social media data, because people post on social media if they're seeing something, you know, before even they call the call center potentially.

So this actually helps us triangulate, you know, the problem zones in a much more efficient manner. So now the principles of managing data still remain the same, which is how do you ensure privacy? It's actually much harder because the documents could contain PII, could contain sensitive data. So you potentially now need to start thinking about putting some basic filters to make sure that you're validating that there is no PII or sensitive data that is moving across into the AI models or into analytics. But the way you

apply technology will be slightly different, but the principles remain the same when it comes to governing that unstructured data.

How do you ensure quality? Unstructured data is a lot, lot more noisy than structured data. just because of the fact that it is mostly human generated data. Even if you're thinking about audio files, there's potentially a lot of background noise, which makes it harder to analyze the voice and the sentiment, et cetera. Could be issues in the video files just by the resolution, et cetera.

The data quality takes a very different perspective when it comes to unstructured data. The good news, again, is that there are technologies available to clean the data in a very efficient way. But net-net, I mean, if you have to summarize, I mean, there needs to be a deliberate strategy towards managing and monetizing the unstructured data.
One part of it is to optimize, which is reduce your cost. You don't want to throw away the data because it could be useful in the future.

The second part of it, which is the, you know, the short, medium term is, you know, work backwards from the objectives you want to achieve. We took a couple of examples, right? Consumer experience, you know, preventing, you know, water losses. There are several, you know, we've even done work in terms of, you know, how to apply AI and the video, you know, technologies to determine the quality of, you know, pizzas that are coming out to grade them. So what is the purpose of that? That's consistency and consumer experience again. So at the end of the day, if the companies are focused on that strategy and then delineating between cost and value, that's a great starting point. And then start applying the standard principles of data governance, privacy, security, et cetera.

Andy Packham
So Srini, just staying with that, because I think that was really interesting about the examples. And, you know, is that where you start in the customer conversation, rather than talking about how I'm going to structure your unstructured data? It's the business value. Is the place to start thinking about, is it business value first?

Srini Kompella
Yes, because you could apply a lot of these technologies and run a lot of experiments and prove a lot of concepts, so to speak. But that's just cost. So that's why it's critical to work backwards from the business objectives, like consumer experience. Now, we took the example of unstructured data here. There's a lot of structured data analytics as well that people have been doing, which can be cross-tapped again. Take that retail experience we talked about.

We can cross-tap that with the loyalty data, for example, customer loyalty data for retail stores, and create potentially newer set of experiences for consumers. right so it's critical to start with the business objective and then work backwards because the good news is there are technologies that are now available compared to a few years ago Jeeva alluded to it you know where you know a few years ago it was much much harder to you know do whatever we are trying to do today today you can do this much faster better and cheaper Jeeva, I'd like to get your views on that about how you're seeing organizations. I mean, you must have this breadth of organizations you're seeing. How are they getting value out of kind of thinking about this unstructured data from a business perspective?

Jeeva AKR

Absolutely. First of all, Srini, that was a very detailed response, very powerful. And if you think about it, I will start with an example, Andy, just to add to what Srini said, and also to provide an additional comprehension for all the viewers.

One, if you think about it, if we were running a Marmon shop, a sunglass shop, so if you have been actually selling 100 sunglasses in January, 200 in February, 500 in March, and then about 1,000 off in April, if you have been just using the structured data that came from the POS systems, And if you were to make a decision on how much of an inventory that you are actually going to order for the summer, definitely keeping in line with the trajectory of the order that you have seen from 100 to 200 to 400 to 1000, you are actually going to stack up your order with several thousands for the summer, anticipating that it is going to be an increase. That's the POS data that coming from the structured systems. Right. Then the second thing is now if you try to combine that structured data with unstructured data, maybe it is that unstructured data in this particular one could be the reviews and the feedbacks that we are actually getting in from the customers who are actually purchasing the product for the last four months time.

And you saw an exponential increase in the dissatisfaction. OK, so people are complaining, complaining about the quality of the product or the fragility of the entire sunglass, whatever the case may be. Right. People are disappointed with the quality of the sunglass that they have been using. And in April, it hit the fever pitch, about 700 people giving a bad review about the sunglass that they bought in the prior three months. Like I said, if you were eliminating that, if you don't include the unstructured data, and if you were to make a decision only based on the structured data, then you would have definitely ordered 2000 or several thousands.

Now that you know that people, this product has got a high dissatisfaction ratio, then you're not going to continue to place an order that is going to be in several thousands probably you will be thinking of pulling the entire rest of the stock out of the shelf right so I'm just giving a very layman example about the power of combining both the structured data and also the unstructured data when it comes to analytics right now you can imagine the power of doing that in real-time analytics with aura with training the iml models and all that so where we see is one Like I said in the beginning itself, that there are two important statistics that I want to share with all the viewers.

One, more than two, the number of companies who are going to make an increased investment in AIML technologies is going to be 63% of the entire world's companies are going to make AIML investments in the next five years time, okay? And then the second thing is that by 2025, which is not far away from now, we expect that we are going to have nearly 200 petabytes of data that is going to come. And we know that 90 percent of it is going to be from unstructured data.

Right. So until four or five years ago, the majority of the problem and the challenges that our customers were actually trying to solve for were related to how can I actually harvest all this data? How can I actually bring it? How can I store it? How can I make use of the data? And that's the challenge that they were dealing with until today. But now the challenge has actually shifted to how do I get value from the data? Because that data acquisition part has been already solved for, right? I think.

Srini alluded to it, both in terms of having the infrastructure capability to ingest all this massive amount of data, and then also how do you make sense out of the data with the quality and all the other stuff. There are technologies that are available. So now the entire world's focus is now, the customer's focus is now shifting to How do I get more value from the data? And from that perspective, I think we have seen increasing amount of technology that is coming in. We have talked about LACOS architecture. We have talked about the evolution of Spark technologies, which is an automated way to bring quality profile for all the data. And Srini alluded to the fact that how do you actually keep the data for persistent storage for a longer period of time? And we have a lot of this technology that has come in the way that is making it possible for customers to solve for it.

Andy Packham
Jeeva, thanks. I think this driving value I think this is an absolute key thing. Do organizations need to think differently about, or does the CIO need to think differently about their IT when they now come to this?
Typically, we've got our high-value ERP systems, and we've protected them, and we've cared for them, and then we've got all of that other content that sits somewhere. I think we're focusing on the high-value ERP.

Jeeva AKR
Absolutely, Andy. I mean, you started this whole session with the business value in mind, right? So the CIO's function is about empowering the businesses that they serve within their organization. So from that standpoint, whether it is about data acquisition or it is about empowering the businesses, it all comes down to what business use cases the company is actually trying to solve for, like the example that I just gave. And also Srini also went through a very real enterprise use case example that we talked about, right?

The power of real-time analytics with all the data that you collect from the machine data and all that. and uh machine data like I said in my opening uh statement itself telemetry data which is actually the data that collects all the interpersonal communications and also the data that is coming from all the devices when I say devices it doesn't mean that it is only to IoT data but also from our own computers they know the click stream data the even logs the security logs all this data that is actually going to provide insight about how we are actually interacting with a particular product or a service which is going to be harvested and it is going to be used for providing a very rich experience for their customers and the users of the product and also for their internal employees.

That's the quest of the digital transformation that we are talking about, right, Andy? So in my opinion, I think it is on the top of the mind for every single CIO today to think about how they can actually you know they have this they solve for all the technology challenges related to harvesting all the data but it is now about getting value out of it and how they can actually create a cycle where they can actually uh you know get insights from their own data they identify the use cases and then they leverage all these different data types to solve for the business use cases that they are thinking about.

Andy Packham
Srini, what's your views on this? Do you see that same pattern? I mean, what should you, you know, those conversations, what are you advising CEOs to think about?

Srini Kompella
Yeah, I think, you know, so to Jeeva's point, right, I think, you know, the data sets need to start coming together more to create the maximum value. Now, what does it mean from an organizational standpoint, right? I mean, you know, if that's the question, then there is definite merit in terms of, you know, creating a data organization combined with AI that is not just looking at structured data, which is what most of the industry has been doing over the last few decades, but also unstructured data. Because at the end of the day, we talked about the data fabric as an example.

You can build data pipelines pretty rapidly today. But you need a cohesive strategy in terms of how you're going to look at data as an asset at an enterprise level. So which includes structured and unstructured. Jeeva said it a couple of times. I mean, there's more unstructured data today in enterprises and in the world than structured. So there is more intelligence and we can create out of unstructured data than potentially structured in the future.

So this could drive the CIOs to start thinking about how to create a more cohesive data organization that looks across structured and unstructured data, number one. And number two, you know, also set up, you know, the processes, you know, because it is people process technology. Like you said, technology is mostly available today to, you know, to solve a lot of problems and create opportunities.

But, you know, the people part of it and how we are, how the CIOs organize the teams. And secondly, the processes. We talked about data governance as a small example. you know, the organizations need to be consistent on governance of data across structured and unstructured. Now that can become simpler if the CIOs actually organize the data teams into one cohesive organization, right? Now, that is a possibility. There is merit in that discussion, but there are obviously, you know, different scenarios that the CIOs need to go through because you said it, right? I mean, the CIOs are, you know, empowering businesses to, you know, apply technology for better results, right? Now, That organization can be structured in different ways, either as a central team, like the way the data organizations for structured data got set up before. It could be in terms of aligning with those specific business objectives we talked about. So let's say we're talking about consumer experience.

Does it make sense to set up an end-to-end team focused on consumer experience? Spanning across AI structured data, where the team is completely are focused on enhancing consumer experience right so there are different permutations combinations but there's definite merit in terms of you know streamlining the way the organizations are looking at data across structured and unstructured and can be solved organizationally you know with some of these approaches .

Andy Packham

We could talk about this for hours and hours. I mean, this has been a super interesting conversation. And I think, you know, for me, the key point is this isn't a technology problem anymore. It's a problem about understanding the value and driving business conversations and business decisions and building strategies around that rather than kind of talking just

about the technology. So Jeeva, Srini, again, I really appreciate both of your time, both of your insights. It's been absolutely amazing.
Thank you very much.

Jeeva AKR
Thank you, Andy. Thank you, Srini.