AI LEADERS PODCAST: META DATA – DATA AS A PRODUCT SHIRSHANKA DAS AND TERESA TUNG

AUDIO TRANSCRIPT

SHIRSHANKA DAS: Technical meta data, business meta data and operational meta data, really can come together beautifully to solve bigger problems.

TERESA TUNG: Hello, everybody. Welcome back to our AI Leaders Podcast. This section is going to be about Meta Data and its role in Data Products, Data Mesh and AI. This is Teresa Tung. I'm Accenture Cloud First Chief Technologist and I'm so happy to be joined by my friend, Shirshanka Das. Shirshanka, could you give us an intro?

SHIRSHANKA DAS: Yeah, absolutely. Hi, Teresa. I'm Shirshanka, CEO of Acryl Data. Prior to this, I spent a decade at LinkedIn as the tech lead for the big data team. I founded a bunch of projects, but most importantly founded DataHub as part of that journey. A big fan of the great work that this podcast is doing and really happy to be here.

TERESA TUNG: Thank you for joining us and Shirshanka is a master of meta data or data about data. So maybe we could start with that. Why is Meta Data the answer to many of these data questions?

SHIRSHANKA DAS: Well, I think I can probably no longer count on my fingers or toes at this point, how many times I get asked this question. What is meta data. I think we should just start there. There's a lot of classical literature on this. But I think for a lot of data practitioners who are just jumping into the workforce and they have very different views on what meta data even is. So let's get right into it.

There are really three kinds of meta data at the top level. The first is called technical meta data. And this generally is all the declared definitions about your data assets. So usually, back in the olden days, it used to be just your schemas and maybe your DDL statements, your create table statements and things like that and your access grants. But fast-forward to now and it includes schemas about everything, your Kafka topics, your operational data stores and the DDL has kind of evolved over the years to becoming not just a DDL that your DBA is deploying to your production system, but maybe in a data Ops world, something that you're checking into a repository that's getting automatically uploaded, like a DBT model or things like that.

And technical meta data is actually getting broader because it's not just about data sets, but also about AI things like models and features. And so, the definition of a model, the definition of a feature and when you look at pipelines, the definition of a DAG, all of these are really declared definitions. Someone, a human in a lot of cases, declared it and checked it in most likely or it's part of kind of your deployment system. So that's technical meta data and that's – you can think of it as the slowly changing dimensions of your meta data warehouse.

On top of this is layered business meta data and this is really the high level conceptualization of your ER model. So at a company like an ecommerce company, you would have users and payments and orders and this and that and you have a web of dependencies across these things and then, you would layer on concepts like a business glossary and other logical concepts and documentation on top of it that allows business users who are not in the daily grind creating tables day in and day out, to get a sense of what kind of data do we have or what kind of models do we have or what kind of features do we have. And that's business meta data, typically operated on by humans who were a little bit away from the nitty gritties of operating on tables and data sets and pipelines.

And the third and very important component of all of this is meta data is operational meta data and this is really the heartbeat of your data ecosystem. Every time a data set gets transformed, a pipeline gets run, someone deployed something to production, all of these active streams of meta data that include like data set landing times, data set profiles, statistics, model training runs, all of this stuff, which is like a firehose of data in a way, but firehose of data about data is what we call operational meta data.

And so, the trifecta of technical meta data, business meta data and operational meta data, really can come together beautifully to solve bigger problems because one of the big problems that people are facing today, too many tools, too much data and I have no idea what's going on in my data ecosystem, but I want to be data driven. And I think meta data is the answer because once you have a good grasp of what is your technical meta data and how is it changing, what is your business meta data and who's writing it, who's producing it and what does your operational meta data look like, which gives you a sense of how your data ecosystem is changing, you can do amazing things.

TERESA TUNG: So in short, it gives you the ability to know what data you have and then, it helps you make the decision of if you can use it and for what?

SHIRSHANKA DAS: Yep, and most importantly, how it's changing.

TERESA TUNG: Because it does change. And so, I think we've done – people have done a difference between meta data and maybe static meta data and what you're describing is something more towards active or living meta data when you talk about the operational or even as you onboard new sets and those sets change in time. So I think that's a big change compared to maybe how we've looked at meta data in the past?

SHIRSHANKA DAS: Yeah, yeah, I think there actually have been always two kinds of approaches to meta data. There was the cataloging approach to meta data where a system or a catalog would come in and say, I'm going to observe things that are happening and kind of give you a clean room experience about your data ecosystem and that was always after the fact. That was always slightly inaccurate, almost like looking at a star very far away and saying that star is very bright, but maybe that star died a year ago and the light just hasn't reached you yet.

And then, there was the other approach which is PTL tools like Informatica took where meta data was part of the tool, where you describe your transformation and as Informatica is transforming your data, it is also capturing meta data and that's just meta data inside a tool. And that, of course, would be active because that meta data is being changed as the tool is transforming it. But what I think we're seeing in the industry is a move towards this sort of active meta data moving out of siloed transformation tools and becoming something that gives you global visibility about how data sets, models, features, pipelines, dashboards, charts, are getting transformed and changed in sort of the same way as earlier we were only looking at technical meta data.

TERESA TUNG: And so, it does bring us to a little bit of the next question. What's the role of meta data in AI? How does good meta data help AI in understanding these data products? I think you talked about a little just now. Maybe you can expand on that?

SHIRSHANKA DAS: Yeah, I think meta data and AI are actually two peas in a pod and they just don't know about it yet. So if you look at where the AI ecosystem is going. We're getting much more formalized about how we're doing AI. We're starting to give models and features their own identity and identity comes with meta data, names, descriptions, context around who trained me, when was I run, what were the parameters used for this run, how much of a needle did I move when I was deployed to production? And this is me model.

And then, there's an entity like a feature that maybe brings in certain inferences attached to it and from a regulatory perspective, we need to know that this maybe is a feature derived off of a sensitive attribute. And those kinds of things are again, meta data about AI assets that we now are realizing is very important to track. So we're starting to move from the kind of early chaotic days of just doing AI and I moved a bunch of metrics and I'm feeling good as a data scientist to becoming – no, no, no, I need to be much more scientific in how I do AI.

And so, that's meta data applied to AI, but then you can flip it and say, well, what is AI's role in having good meta data itself? And one of the things we talked about is how meta data itself is kind of like a big data problem because we've moved from just the technical meta data or even the business meta data. I mean if you go into organizations and ask them how many data sets do they have?

Even at the largest of organizations, like about a million plus give or take. That's still a lot, but still it's a least in the millions, we can still fit on a single NoSQL if you've got a nice box.

But when you start looking at operational meta data and you start looking at every single type of pipeline runs all of the data that it produces and you add all of that up, well, now you're talking data lake or a big data implementation. And the moment you get there, now you start running into the problem of how do I make sense of it all? Because if you start connecting up your operational meta data streams, your technical meta data and your business meta data into a single graph, which is basically what we do with DataHub, you start needing to answer the question, I typed in a query like customer and I get back 20,000 results. So which one is really the customer data set? And that's where AI comes in. AI comes in by saying, well, I'm going to help you get signal out of all of this high-fidelity noise that you've collected in one comprehensive graph and I think there's a lot of space there for AI to help in making meta data delightful, easy, relevant, consumable and good. Because there's only – and we can get into that later – there's only so much you can do with human powered meta data efforts.

TERESA TUNG: You're going need both. So meta data will help AI by helping us find the data that we want to use and checking that there's quality, there's lineage, governance. Also, the models that are part of that data product and then, you're saying the other way as well, AI is going to help meta data because there's no way that we're going to track, as humans, all of the places that we've just democratized AI using meta data. We've just automated AI into production using meta data. And so, you're going to now need AI to actually manage that because there's no way to manage it otherwise.

SHIRSHANKA DAS: Yes, AI and meta data needs to be BFFs for a while.

TERESA TUNG: So let's talk about the latest hottest topic, Data Mesh. And I think it's aligned with everything we spoke about so far. Data as products. AI also into production democratizing that use. So Data Mesh certainly is around data products, around self-service infrastructure, decentralized domain ownership with the federated governance. And so, this term that Jamak Degoney (ph) coined, I think there's been a lot of rallying around it, I think any data nerd reads it and resonates as something that is a vision that we're all working towards. I think we've been working towards it as an industry.

But do you have some concrete examples of Data Mesh and specifically maybe how Meta Data is really important to realizing this in industry?

SHIRSHANKA DAS: Yeah, I think Data Mesh, as you rightly pointed out, is kind of a combination of something we're talking about now and also principles that we have been trying to accomplish as a data industry, although we probably never bundled them all together under the current Data Mesh.

TERESA TUNG: That's the greatest thing she did is to bundle it together and rally us.

SHIRSHANKA DAS: Yeah, so that's been the greatest success of Data Mesh. I mean not to put a fine point on it, but one of the things religion does it is put practices together and it says, here are the five things you do as an X and that makes you an X, whatever that X is. And so, I think Data Mesh has done something similar. And so, it also suffers from some of the same things that religion suffers from which is everyone has their own interpretation of what it means.

You can take the principles of a thing and then implement it in your personal life in a slightly different way. And so, data practitioners, I think, are taking those principles and then applying it into their practices and their organizations. And one of the things I can definitely say I'm fortunate is because DataHub, the open source project, is not just an open source project, it's a community of data practitioners. So I get to meet a lot of – we get to meet a lot of companies that are implementing DataHub at their organizations. And I get to understand how they are implementing data practices at their companies.

TERESA TUNG: Maybe could you just pause and talk about what DataHub is? I think we've mentioned it a few times, but maybe just interject right now since you mentioned, what is DataHub?

SHIRSHANKA DAS: Yeah, DataHub is a Meta Data platform that is built for the modern data stack. We are taking a developer first approach to storing, managing and visualizing meta data and we get – so the way most teams discover DataHub is starting with like a data discovery problem and moving onto a data governance problem and, in many cases, also moving onto a data observability problem. So we see people, data platform teams, data governance teams, data analysts coming into the community with kind of one of these three burning problems.

One is I cannot see where my data lives and how it's getting transformed and where it's going. And related to that, I don't know how to govern my data well in a scalable, sustainable way. And thirdly, my data is changing too fast on me and I can't keep track of what is breaking where and what I need to be paying attention to. So those are kind of the three big use cases that people come to DataHub for.

What is it? It's a project that I started at LinkedIn. We open sourced it early 2020 and it's been a year since I founded Acryl Data to commercialize the DataHub project and really take it to every single enterprise that can have a better data culture and have a better data experience through a product like DataHub and it's been a great journey.

So coming back to the question that we were addressing, which is what are concrete examples of Data Mesh being realized in the industry? We have a company called Saxo Bank. They were actually were an early adopter of DataHub and also an early adopter of Data Mesh. And one of the interesting things they have done with DataHub and Data Mesh is made DataHub sort of like the control plane of their Data Mesh implementation. So what does that mean in practice?

Their entire business glossary, which is a very old school meta data term, a business glossary is your classification terms, your customer model, your partners model, whatever your business terms are, they've converted all of that into protocol schemas. And then, in their technical schemas, which are maybe Kafka topic schemas or table schemas, they import those business glossary elements and they annotate their technical schemas with their business schemas.

These schemas get checked into their DataHub repositories and it's a completely federated model. So you could have Team A owning and managing their own schemas with connectivity into the global business glossary. Team B, managing their own technical schemas with connectivity again to the global business glossary. And as these schemas are checked in and deployed to production, the CIDC systems are checking for backward compatibility, but also dropping in events on the meta data bus, if you will, or the meta data highway, depending on whether you like small things or big things.

But this meta data stream is coming into DataHub and so forth, everyone else who is either a producer or a consumer of these data products, they're able to go to DataHub, find the data product, they're able to find who the owner of a data set is, they're able to find other attributes about the data set. They're able to see how it's versioned over time, pretty much the concrete implementation off a data product with it's meta data storage in Data Hub as a backend, as well as its discovery and management experience in DataHub as a front end. So DataHub is two things. It's the Meta Data platform, as well as a product on top that gives you kind of the visualization on top of the Meta Data.

TERESA TUNG: And so, this is again another example where Meta Data is really the answer that stitches all these different domain owners and consumers together. So you mentioned that these different domain owners can produce and publish their own products that the Meta Data Hub dropping it in the bus or the highway and at whatever speed makes sense for that product. And then, as consumers, we subscribe to the ones that we're using and maybe part of the guarantee of the product is how the product is now and I need to know, so that I can make a change or if the product's going to have a versioning change or new feature. All of this is exactly what it means to go from a data set or a feature store to a data product.

I think without this, that product notion where I can base my business decision around your technology solution, I think that would be missing, right?

SHIRSHANKA DAS: Absolutely. And one of the things, for example, they are integrating right now is great expectations. So I don't know if you're familiar with great expectations. It's a pretty popular open source tool for data quality. And so, Saxo Bank has implemented data assertions using Great Expectations. And so, what they're doing is piping the output of those Great Expectations runs back into the DataHub console, so that you essentially have a one stop experience, kind of like a control center for data where you can come in, you look at the definitional side of the data set, which is coming in from the CICD systems, but you're also able to see the operational help of the data set in terms of data quality scores, data quality assertions that as a consumer, as well as a producer gives you much higher confidence about is this a data set I'm going to bet my next promotion on or my next project on.

TERESA TUNG: And how you differentiate the product to begin with, so this gives you the business case to say, I'm going to have a better data product because of quality with great expectations or I'm going to have a better data product maybe based off of the freshness or the easily integratable, the format thing. So I think you just gave some really great concrete examples of your work with this thing, both with the data Ops process, it was completely federated, different people were able to publish into the bus and then, also with what really makes a data product great.

I did have a question about that role of the centralized organization and decentralized. And then, what you saw here, is it the roles – who sets up the architecture, right. I think the Meta Data Hub or the DataHub still is the central org, right, and then, is there anything else that that central org should be doing beyond the technology?

SHIRSHANKA DAS: Yeah, that's a great question. In fact, again, we hear this a lot in our community where people want to go on this journey with us and they're unsure if it's a boil the ocean kind of journey or it's an incremental journey and who is really in control and who needs to buy in and what should the organizational structure be to actually have a successful outcome.

So what we've seen repeatedly in terms of success stories is when you have a central – let's call it a platform team, who is notionally charged with the responsibility of owning and operating the data IT function. So they may not run each one of these systems, so for example, in some cases, people are just working with Acryl Data and we host DataHub for them, but at least they are responsible for making that decision, like what central Meta Data platform should be used? What BI tool should be the offering to the stakeholder and things like that? But then, they often have very demanding stakeholders. Stakeholders like – and this was basically my whole past life, so stakeholders tend to be like the CISO, who is barking orders at them about security, security, security and compliance, compliance, compliance. They also have the data governance team, who sometimes is slightly different.

They're all about make sure data hygiene is being maintained, make sure we're being conscientious about our user data and this kind of emerging data privacy practices. And then, there's the data engineering team, who sometimes is – and the data science team, who are kind of sometimes either the same team or just very closely adjacent and they are primarily impacted by upstream teams breaking them constantly all the time and they're just always running around trying to fix a data pipeline that got broken or a data lake data set that's not of high quality and trying to fix it up.

And then you have the ML practitioners and the AI kind of world who's also asking the data platform team continuously for better insights, better infrastructure for them to produce models quicker and faster and ship product quicker. And so, what we're seeing repeatedly is these central data platform teams are making platform decisions about how should these different – how can I scale myself effectively, how can I bring in a tool or a technology that can address all of these disparate questions that come to me, all of these disparate requests that come to me, while not needing to grow to a 200 strong data engineering or data platform team.

And so, this tool choice that they're making, often tends to be what is the most future proof tool? What is the most future proof technology that I can bring in that allows me the flexibility to evolve as the tool stack evolves. So I talked about Great Expectations as a data quality tool, but there's actually like a lot out there. And even as we were talking probably another data quality startup got funded.

And take that with data governance, take BI, like all of these evergreen spaces where innovation continues to happen, central teams are constantly having to make decisions about what tool to bet on, but they know that these decisions are not going to be super long term. So they're always looking for technology and tools at the lower level, the substrate, if you will, that can allow them to make changes later on without having to rewrite the whole stack.

At the same time, there is the top down decision around I want to go Data Mesh and what does that mean? And what we've seen repeatedly is there is typically a CDO or a CIO or someone who basically says, this makes sense for us. Why are we not doing it? Let's make it happen. But going from there to actually having a successful end-to-end implementation, typically requires partnership with a producer and a consumer team. So generally, we see the data platform team having a relationship with a data producer team, maybe the most impacted one and a data consumer team maybe again the most impacted one, and building a reference implementation for what an end-to-end Data Mesh might look like for a single domain. And that's where practices, being able to create central policies that can then we disseminated into these CICD pipelines into how things are getting checked in. Those are opportunities to implement the first version of the Data Mesh for this company.

TERESA TUNG: So it is very much like the Mesh promise, right? It is data product led and so, we're going to lead with the use and the outcome and whether it's from a consuming team and the product they desire or from a domain owner and a product that is truly useful for their business and across the business. I think that that's a good rule of thumb of how do you get started.

And then, using that to get the right technology into place and really demonstrate, like any platform, any platform also owns the killer app. So I guess, you're saying that the same sort of journey that you're seeing in this space as well?

SHIRSHANKA DAS: Yeah, and we often see – you're absolutely right, we often see visibility as one of the number one kind of drivers of that first use case for like a Data Mesh implementation. Another one is often data quality. One thing I did not talk about that is actually a very important hidden figure, I guess, of a successful Data Mesh implementation is good data modeling because you cannot tool that out of existence. You cannot federate that out of existence. I think a good consistent data model requires some amount of holistic part.

TERESA TUNG: Oh, so human thought and human agreement across teams, is that a more outspoken way of saying it?

SHIRSHANKA DAS: Yeah, yeah, and I think the one thing that people should not assume is that they can just federate everything out and then, magically an amazing customer model will emerge that they can do analysis on.

TERESA TUNG: Because then, none of the data connects, none of it is easy to use. So everything against a product is that typically a product lines like the LEGO sets that connect together, building blocks, right, that the model adds to the other set. It's a bonus and without that, it sounds like you just have a lot of disparate pieces that don't work together.

SHIRSHANKA DAS: Yeah, and I think that's where our central data team and maybe an adjacent central team can help with being the tastemakers for what does a good data model look like at this company? How do we name or entities? What are these IDs going to look like? And let's make sure we name them consistently, so that the joins work. If we are going to be of any denormalized data model end-to-end, let's all be denormalized. But if you're going to be very normal lifestyle, let's all be normalized, so that when someone's consuming data products from two different domains, they don't feel like they just don't feel like designed by a similar - in similar taste. And I think that's where there's some interesting challenges left for our data industry.

TERESA TUNG: And so, you talked about that pivot, right? So as you move towards this federated organization, the central team, in addition to the tools, right? So as part of the tools, there might be patterns for how you push meta data or how you if you use the tools already, if it's Informatica, in your example before, I've already connected, how to automatically get that active in live meta data. If I don't, here's a design authority for how do you do the technical pattern? But just as important, you're saying as the design authority and the design pattern and the data models themselves and then maybe and the AI models and the right abstraction interfaces to make these fit together?

SHIRSHANKA DAS: Yeah. The more you federate, design patterns become even more important.

TERESA TUNG: Right. And so, their role just becomes more important. And this is what's needed to democratize and enable the rest of the business to try all the weird stuff or to keep the old legacy thing up and running, or even if it's all in the same data lake physically, this allows each team to be able to evolve their own data product.

SHIRSHANKA DAS: Yes. So the best central teams are going to be the ones that can transition from being doers, which is what they are today, to being enablers and shepherds and being able to get out of the way where it matters while still preserving consistency.

TERESA TUNG: So we already touched a little bit about DataHub and how it came from open source. I want to talk about, you know, what is the role of the community in shaping the sort of modern data catalog?

SHIRSHANKA DAS: Yeah, opensource is a very interesting thing because itself or what Open-Source means and does has actually changed over the years in software. Back in the day, Open-Source was a way for libraries to get shared across companies, right? And so, there's a lot of like really important open source libraries like open SSL.

Like it's probably everywhere. JTBC, it's probably everywhere. So this is a huge like I don't think anyone can contest how important open source has been to making sure that the foundations on top of what we are building is like super solid. And the reason why that has happened is because those libraries in the past have been the most thoroughly tested because it's had the most amount of eyes on it, the most amount of even hands on it, because you might have had a lot of contributors writing code in it and improving it one bit at a time. And just over the years, it just ends up becoming the most battle tested, solid piece of technology, but it takes time to get there. In the beginning, it's kind of like an awkward teenager and not quite there, right? It's got some, you know, facial hair, but, you know, scrawny, and that's kind of that awkward growth phase that open source projects go through. And we see the same thing emerging in kind of the product space.

If you see, for example, how DBT has taken off among the data engineering community as an Open-Source tool, it gives people accessibility to something that would otherwise be locked up behind a vendor demo wall or a pay wall. And you know, the inertia is so high to click through and finally get access to something and then try it out. Then most people just don't try something like this. So the thing that we're seeing with DataHub itself, it obviously got created at LinkedIn when I was there and when we felt like we had something that was useful and shareable with the community, we open sourced it.

But then I think the year of 2020, as I was shepherding the community and trying to help them get successful with DataHub, I noticed that it's one thing to open source a project that was built at a big company and it's quite another thing to make it successful at tons and tons of enterprises that are maybe running slightly different stacks, have very different kind of tools that they want to integrate it with. And so, when I founded Acryl Data, the whole mission really was, I want to make it super easy, super delightful and super productive for people to take DataHub, the product and then make it change how they do data at their company.

And so, we're seeing this emerging modern data stack with tools like DBT, Great Expectations, DataHub and some of the ML stack, like Feast things coming along where people are putting these things together and saying this feels like a consistent, modern way to do data, and the one thing that I had not realized was the power of the community. So open source in the past was more about the code is out there. You can take a look at it. You can send some contributions over. We'll have some conversations over email, right? Or maybe there was an IRC group, but not very common, right? But the modern discourse is really different. We have a slack community of almost 2,000 data professionals right now. When we started the company earlier this year, it was just 200 and something. So we've grown almost 10X in just one year. And pretty much everyone that's joining the community is coming there with an intent to get productive and to take DataHub and make something good out of it.

And as part of that community, we are having discussions that are not just about the project, that communities are having discussions like how should we do ownership? It's a huge hard problem and I don't think one particular tool can just come in and say, oh, I got ownership figured out. It's just this little button over here that says, add owner and, you know, you go click that button and boom, now you have ownership. The reality is that ownership is such a multilayered problem.

And so, what we ended up doing was we did a community sprint almost where we surveyed people in the community, shared best practices across each other. And I think this kind of merger of open source with community led, creates a very different kind of product because your code is out there. But not only is your code out there and people are contributing to it, but even your product development phase is out there and inclusive. So it's not like you're sitting quietly writing code and then shipping it to your customers and saying, do you like it or not? And maybe you've got two or three customers that you're working closely with, but you've actually got a thousand plus community members who are actively giving you feedback before even you develop the product.

So I feel really fortunate as Acryl Data that I have access to this large group of people who are giving me constant pre-product feedback about what should an ownership experience look like in DataHub? So that when we actually build that first version of ownership in the product, it's going to be a new way to do ownership that is actually sourced from best practices that individual teams have built in-house their organizations. And that, I think, is a game changer in terms of how modern companies should be operating.

TERESA TUNG: I mean it goes along with how data is used too. So there's a lack of data. We don't have enough data as individual companies maybe, right? But coming together, there's a lot more data cooperatives and collaboration models and often times the AI model that we use, we didn’t make. We're going to download the data sets that we might use or a third party. We might be able to see a better view of what's actually happening by working with our partners in a secure way.

So naturally, the problems that exist are require collaboration. But you can't do it within a line of business and you can't do it even within a single company. So open source seems like, you know, when we're talking about how do you create shared IP where nobody's legal team gets them into trouble, even all the way down to ontologies and data models, right, open sourcing those, so that we can all work and reuse to publish data sets in ways that can connect right or talk about the same concepts seems like the right thing to be doing.

SHIRSHANKA DAS: Yeah. And I think you've touched upon two - one interesting thing, common data models, and that's actually showing up in two different areas. One is common ontologies, like you said, what does customer look like, what does orders look like, what, you know? So that's one. And we're seeing the community starting to talk about contributing some of those ontologies back into the DataHub project, which I think is going to be amazing.

The second thing that you also talked about is common meta data models, and that's also interesting. When we are working with, for example, the orchestration providers in the modern data stack like Airflow, Dexter, Prefect. We're starting to see an emergent standardization of what does meta data model for a pipeline look like. And I think for some of these things, it is premature to do the modeling before doing the input, like all. You know, if you think too much about the data modeling, you are going to get it wrong. You actually want to just get in, do your first version, realize that it was wrong when you do the second one and then kind of iterate. I am fully in favor of the kind of iterative approach with eyes wide open, obviously, approach to data modeling. And I think similarly, we should take the same approach towards meta data modeling. And I think we're seeing kind of that happen with DataHub where as we bring in more and more tools into our fold in terms of integrations that we provide, we're starting to see the meta data model get broader and more interesting. And we're starting to support things like generic meta data models like everything is a data set, but some data sets are also streams, and some data sets are also views and some data sets are actually Kafka topics. And that's very specific.

But you know, you need to have a model that can do the generic stuff well, but also allow for like infinite extensibility and infinite specialization for you to truly capture the specificity that lies inside each of these tools. And we've seen kind of a lot of that develop as part of these community driven conversations and integrations with kind of the extremely large set of tools that we integrate with.

TERESA TUNG: And that brings us to the last question where we wanted to land, like what are the characteristics of a good solution for meta data for Data Mesh implementations?

SHIRSHANKA DAS: Well, I think it's probably not just for Data Mesh implementation, but I think Data Mesh does make certain things much more front and center than maybe in the past where Data Mesh wasn't a concern. And I have written a blog post about this, I think it's been like maybe to this day, pretty much a year since I wrote it, and I try to write about the evolution of Meta Data architectures as lived through my personal journey in designing and evolving the DataHub product. Teresa, you might remember warehouse from Meta Data.

TERESA TUNG: I remember I was about to say it's a great blog on LinkedIn, right? And you talk about that warehouse and I'll let you describe it. But it's a great blog that shows how that this journey, right?

SHIRSHANKA DAS: You know, when we started, we were bright eyed, bushy tailed, we didn't know what meta data really meant. We were just building a search and discovery product to make our data analyst productive. And so, we did a Gen One version of the architecture, which is connect up to everything, pull everything together, crawl, crawl,

crawl, pass, reverse engineer, try to figure something out somehow and then build a UI that at least makes you find things quickly, makes you somewhat productive. And it's not bad. I mean, it's a good product, and it did survive, even open sourced it. But then it ran into trust issues over the years because each of these individual crawlers that we had written, that was kind of reverse engineering stuff from the back side of Informatica and this and that and parsing big logs and we did some bad stuff. It ended up just - now we have 20 fragile pipelines that each individually could be broken on every possible way. And now we have a meta data trust problem. So we've gone from data not being trustworthy to meta data itself, not being trustworthy, which I would argue is a worse problem to have.

Because you search for a table –

TERESA TUNG: It wasn't very domain oriented, right? Because that central team did the pool and you maintained all the pipelines. So if I had a new tool, would I go to your central team and say, please make me that meta data crawler?

SHIRSHANKA DAS: That's right. That's right. That's right. And it's exactly that same anti-pattern that we talked about earlier with respect to Data Mesh. You want central teams to not be as much of a doer and be more of an enabler and a shepherd. So the central team has to say, here's the contract, you, new tool, need to produce to the contract. And as DataHub evolved, it in its meta data architecture started supporting that concept of, you know, push based architecture, which means and a schema oriented model first architecture, which means there is a concept of a data set. It's got an identifier and there are certain attributes, we call them aspects, attached to a dataset that you can produce too.

So if you are a schema deployment system and you're deploying a Kafka topic schema to production, you can tell DataHub about it using a very specific event that you can publish to DataHub. Now it so happens that we use Kafka for this, but that's not the point. The point is there is an API, it's strongly typed and it allows you to communicate a change that just happened in your domain. And when we flipped to that model, it just opened up a whole new world of connectivity to companies that we did not even know existed. Like, you know, even LinkedIn actually acquires companies. And as we see in our DataHub community, there's a lot of companies that are adopting DataHub that are actually companies of companies. And it's a phenomenally easy way to connect up meta data across multiple disparate data silos that you have because each individual data silo can just publish data and meta data that it has about itself onto a common meta data bus.

TERESA TUNG: And I can add unique -

SHIRSHANKA DAS: Because, of course, happens to be –

TERESA TUNG: I can add unique meta data for from my purpose to differentiate my product then too. So again, it almost seems like DataHub is a Data Mesh type approach to meta data, whereas, warehouse might have been a traditional centralized approach.

SHIRSHANKA DAS: Exactly, exactly. It was literally a meta data warehouse.

TERESA TUNG: Yes. Warehouse and warehouse.

SHIRSHANKA DAS: Right, right. It was actually pulling in everything and a central team was becoming the meta data warehouse team and it was just a bad time for them when they did that. And then we were able to switch to a more Data Mesh model of, you know, Oracle Team, DBA team, you produce metadata out. It just led to much better contracts end-to-end and it led to much better pipelines for Meta Data itself. And so, what we're seeing in both the open source community, as well as in our commercial product, is that companies are able to connect up disparate data sources, disparate data domains much more easily. And with the ability to annotate individual sections of data as being owned by a particular domain, they can easily come in and only look at stuff that they care about. So you can have either your blinders on and say, I only care about my domain’s data. We're actually working on that feature for the open source project right now. But we think, you know, this is one of the differentiating features, both in terms of the meta data storage layer, but also in terms of the meta data experience. You can either put your blinders on and be only domain focused in how you're exploring the graph, or you can take the blinders off and kind of get a bird's eye view of cross domain, what's going on, how are things changing?

And analytics, in my opinion, is going to change the game for all of this because as a decision maker, I want to know where to spend my next dollars, right? And so, a lot of times data leaders are nervous about starting something because they don't know when it'll end. So, for example, if I want to uplift documentation. I don't know where to start. But if I have all of my meta data and my Data Mesh connected up and I have got these operational signals, then I can get meta data analytics that tells me which are my top used data sets, which are my most influential data sets, so that I can focus my attention on the data sets that matter instead of running five year long initiatives where it's never getting done and I ended up documenting the data sets that are obsolete by now.

So we're hoping to change how data gets done by taking our meta data driven, metrics driven approach to doing data. And I think it's a pretty exciting time.

TERESA TUNG: Well, thank you, Shirshanka. So clearly, Meta Data or data about data is really what's needed to maximize your data products, it’s critical for unlocking Data Mesh, unlocking AI and I really want to thank Shirshanka for sharing your experiences with us.

SHIRSHANKA DAS: Thanks a lot, Teresa. This was a great conversation as usual and looking forward to hearing feedback from the listeners. Thank you.

Copyright © 2023 Accenture
All rights reserved.

Accenture, its logo, and High
Performance Delivered are
trademarks of Accenture.