Transcript provided as a free resource by 

Data Mesh Radio Episode #113: Data Governance In Action: What Does Good Governance Look Like in Data Mesh

Interview with Shawn Kyzer and Gustavo Drachenberg

Listen (link)

Transcript provided as a free community resource by Starburst. To check out more Starburst-compiled resources about Data Mesh, please check here: https://www.starburst.io/info/data-mesh-resource-center?utm_source=DataMeshRadio

0:00:00 Scott Hirleman

The following is a message from George Trujillo, a data strategist at DataStax. As a reminder, DataStax is the only financial sponsor of Data Mesh Radio, in the Data Mesh Learning Community at this time. I work with George and I would highly recommend speaking with him, it's always a fun conversation.

0:00:18 George Trujillo

One of the key value propositions of data mesh is empowering lines of business to innovate with data. So it's been really exciting for me personally, to see data mesh in practice and how it's maturing. This is a significant organizational transformation, so it must be well understood. Empowering developers, analysts, and data scientists with downstream data has been part of my personal data journey that reemphasized the importance of reducing complexity in real-time data ecosystems, and the criticality of picking the right real time data technology stack. I'm always open and welcome the opportunity to share experiences and ideas around executing a data mesh strategy. Feel free to email or connect with me on LinkedIn if you'd like to talk about real time data ecosystems, data management strategies, or data mesh. My contact information can be found in the notes below. Thank you.
LinkedIn: https://www.linkedin.com/in/georgetrujillo/
Email: george.trujillo@datastax.com

0:01:11 Scott Hirleman

A written transcript of this episode is provided by Starburst. For more information, you can see the show notes.

0:01:18 Adrian Estala

Welcome to Data Mesh Radio, with your host, Scott Hirleman, sponsored by Starburst. This is Adrian Estala, VP of Data Mesh Consulting Services at Starburst and host of Data Mesh TV. Starburst is the leading sponsor for Trino, the open source project, and Zhamak's Data Mesh book, *Delivering Data Driven Value At Scale*. To claim your free book, head over to starburst.io.

0:01:50 Scott Hirleman
Data Mesh Radio, a part of the Data as a Product Podcast Network, is a free community resource provided by DataStax. Data Mesh Radio is produced and hosted by Scott Hirleman, a co-founder of the Data Mesh Learning Community. This podcast is designed to help you get up to speed on a number of Data Mesh related topics, hopefully you find it useful.

Bottom line up front, what are you going to hear about and learn about in this episode? I interviewed Shawn Kyzer, Principal Data Engineer, and Gustavo Drachenberg, Delivery Lead, at Thoughtworks. Both have worked on multiple Data Mesh engagements, including with Glovo starting two plus years ago. So some key takeaways and thoughts from Gustavo and Shawn's point of view. Number one, it's very easy for centralized governance to become a bottleneck. Make sure any central governance team or board that is making decisions has a way to quickly work through backlog through good delegation. Not every decision needs deep scrutiny from top management. Number two, to do federated governance right you need to enable the enforcement or often more appropriately, the application of policies through the platform wherever possible, take the burden off the engineers to comply with your governance standards and requirements.

Number three, domains should have the freedom to apply policies to their data products in a way that best benefits the data product consumers. So if there are data quality standard policies, the data products should adhere to the standard for measuring something like completeness as an aspect of data quality, but their data product might be optimized for something other than completeness when you think about data quality. Number four, the cost of getting anything "wrong" in data has been quite high because of how rigid things have been. The cost of change was high, but with Data Mesh, we are looking to and finding new ways to lower the cost of change in data. So it is okay to start with policies that aren't complete and will evolve as you move along. You wanna kinda think about security and specific compliance things, but especially when you think about the value add policies, you don't have to get them full, 100% where they're gonna be, five years from now on day one. Number five, if you have an existing centralized governance board, that will sometimes make moving to federated governance challenging at best. So you will need a top down mandate to reshape that governance board, look to meet the necessary representation as well across your capabilities. So like product, security, platform engineering, but to look to not create a political issue if possible.

Number six, look to add incremental value through each governance policy and look to iterate quickly on policy decisions where you can. Create a feedback loop on your policies to iterate and adjust, it's okay to not get your policies perfect the first time, you can adjust them. Number seven, really figure out what you are trying to prove

out in your initial proof of value or concept. If it's full Data Mesh capabilities, that can easily take four to six months according to Gustavo and Shawn. An interesting incremental insight from this conversation, Zhamak has warned about organizations trying to scale too fast as an antipattern that may result in lots of tech debt or even a failure of your implementation. Another interesting incremental insight. In all of the Data Mesh implementations, Gustavo and Shawn have worked on thus far, the initial data product has not had any PII as adding PII adds significant complications, probably beyond what the value add of including that PII would be in most cases, when you're first figuring out your Data Mesh platform when you're first building it out. Number 10, your Data Mesh implementation team should be one to two people from every necessary capability. Talked a little bit about the capabilities in number five.

Number 11, Data Mesh is a large commitment: resources, time focus, etc., so you need to be prepared to fund it for the long haul. This isn't an initial big bang approach, but this is also why you should keep focus on continuous incremental value delivery once you get to delivering your data products to keep up the momentum. Number 12 and finally, you will get things wrong as you move forward with your Data Mesh implementation. Look to limit the blast radius. But it's absolutely fine and expected that you will learn and improve. Data Mesh gives people flexibility and flexibility allows for making changes. Set up fast feedback loops and look to iterate rather than trying to get it perfect the first time, perfect is the enemy of done. With that bottom line up front done, let's jump into the interview.

Okay, very, very excited for today's episode, I've got Shawn Kyzer, who is the Principal Data Engineer at Thoughtworks, and I've gotten Gustavo Drachenberg who is the Delivery Lead at Thoughtworks, and they've worked on a large Data Mesh implementation with Thoughtworks. I think they can probably give the customer name since they've been very public about working with Thoughtworks on this. But what we're gonna be talking about is kind of WTF is federated governance? How does that work compared to what is centralized versus decentralized, and then what is federated actually mean? And is it the same word or it's just a different phrasing for it, or the same meaning, which I don't think it is, but I'm excited to dig into what they've done and what they've worked with on a client of taking their governance from not quite as mature to more mature around Data Mesh and what they've learned around that, that we can apply to a lot of other people's implementation. So, there's a lot that we're gonna dig into and we're gonna kinda bounce around within this concept around governance, and I'm excited about that. But before we jump into that, if you don't mind, Shawn and Gustavo, if you could give people a bit of an introduction to yourself and then we can jump into the conversation at hand.

0:09:24 Shawn Kyzer

Yeah, super happy to do so. Yeah, as Scott mentioned, I'm the Principal Data Engineer here in Thoughtworks, Spain, in sunny Barcelona. And yes, one of the first Data Mesh implementations we did was with Glovo, and we continue to work with them. I think we consider this one of our major success stories, they are doing an excellent job. But since then, we've also worked on several other projects as well, and specifically in the area of federated governance. Another thing that I personally am very involved in is the machine learning and data science community, and also like how that plays inside the Data Mesh. So I'll actually be doing a podcast YouTube with the MLOps next week on scaling Data Mesh with machine learning, so pretty excited to get involved in that.

0:10:15 Gustavo Drachenberg
Yeah, my name is Gustavo, thank you Scott. I'm a Delivery Lead at Thoughtworks. I fulfill the roles of product manager, project manager, whatever it needs to get done, I help. In the past, I was involved in a lot of cloud migration, decomposing big monoliths, using domain modeling and using domain driven design into the cloud and microservices. And then Data Mesh came up and we got involved with applying Data Mesh in practice, back in the day when, I'm talking about maybe two years, we just had some articles of the blog post. And from there we had to implement with the clients because they were really eager to get it going, something that resonated with them. And so creating decentralized architectures and Data Mesh, really are pretty much very similar, a lot of things in common. And that's how we've been involved probably since we started the engagement, Glovo and also through other clients in this Data Mesh journey. And helping clients go through this process of like, how do we do Data Mesh in our organization from scratch?

0:11:32 Scott Hirleman
Yeah. And it's funny with MLOps, there's also a concept of ML loops. I wanna kind of get to that around Data Mesh as well, like, what are some of the anti-patterns and some bad pathways that people have gone down, but we probably won't cover that too much in this. But I kinda wanna do one of those black silhouette things where people with the voice change type of things where people can be really honest. Because the people who were doing it, it's still bleeding edge, but the people who are doing it were very, very bleeding edge when you were looking at it a couple of years ago and trying to figure out how does this work. There are a lot of really interesting stories there, but I think let's start with the kind of big question of what does federated governance actually mean? Does it just mean decentralized? Does that mean... How does that concept start to play out and then we can kinda jump into what that actually means from the implementation side with Glovo and other clients about what you've learned around that.

0:12:41 Gustavo Drachenberg

Yeah. Maybe I can add my two cents and then Shawn can help me out. I mean we've got different options when it comes to data governance. You can have no governance, maybe centralized governance, decentralized governance, and then we have federated computational governance with this which is the latest version. And what we've seen coming into many companies is that there's little governance or decentralized governance, which is mostly informal. It's the best attempt to provide some policies and some guidelines. But it's very hard to think and just make all these pieces work. And then you have a kind of centralized governance, which you have a very strong board, which has a top down that every policies that has to be enforced and very, very rigid. And the other side we have federated computational governance which we think it's something more healthy in the sense... For instance we don't call the data governance board, a board. We call it a data governance team because it's a team that facilitates some policies. And then the part of computational is where the platform comes in to provide some tools so that the product teams that are building Data Mesh products can comply with the policies that the governance board has recommended to keep the company safe, legal compliant and out of trouble.

0:14:18 Shawn Kyzer
Yeah. And I would just kind of add to that. Just a little bit of background. Gustavo and I, he's kind of the yin to my yang that is to say, he's very product oriented and operational, like with the people and the processes. And I'm very much kind of thinking about, okay, we have this, how do we implement this from a technical perspective? How do we automate these, these policies? And so when I think of federated governance, there's the classic definition, which Gustavo mentioned, but I also kind of think of a bit of a metaphor, right? Like we live in a free society, but you know, we can't, you know, break into someone's home, right. So there are some guidelines that we need to abide by, but like with the federated data governance, it's not as specific, right? You have a lot more freedom.

So for example, with something like data quality, you may say, okay, there's a policy. And in fact, every data product must have data quality in these four different areas, right? Accuracy, uniqueness, whatever the organization decides. However, we federate out how those metrics are measured, what the formula is, right? That's the responsibility of the domains. So there is some guidance, but it's very different from the traditional sense of governance where it's super specific, like master data management, for example, right. Like we don't really try to do anything like that in the federated, data governance space. And then yes, there's the computational aspect where as much as possible that we can automate within the stack at the platform or at the data product level, we absolutely 100% do. We want the policies to exist and the automation to exist. And then we want the data, product developers and the platform folks to not have to really think about it. We just want it to happen as part of the self service infrastructure.

0:16:18 Gustavo Drachenberg
Yeah. I was just gonna say, if we want a practical example, let's say that the governance team will say that your data products need to provide metrics around their data quality. And they might recommend some dimensions that can be measured and then the platform will provide the ability of measuring those dimensions. And then the teams that are implementing the data products will use those tools to provide maybe the metrics that apply better to the data that they're handling and comply with the policy that their data products need to have observable data quality.

0:17:00 Scott Hirleman
And I think another aspect that I don't think we need to get into today is the computational aspect of like that there is governance around how things are actually created and processed and things like that, which Zhamak has talked about, but the tooling around like really good cost controls and things like that and measuring that. I kind of feel like governance is such a bad word for what governance should be, because one it's like it feels like it's, you know, heavy handed kind of governance, but it's also, there's like 50 subtopics. It doesn't make any sense, that it's all one thing. But so let's talk a little bit about... So if I could sum up a little bit of what you're saying, so decentralized would be, every domain just controls their own governance. Federated is that it's kind of like a lot of government structures, where you have centralized rule, you have centralized... Like in the US like the federal government, right? Like it's actually having these centralized policies and you have centralized kinds of rules and things. And, maybe not in the US, but in functioning countries, you have a good type of infrastructure that they provide. And then the states themselves can also govern themselves and actually implement a lot of the things.

And that there's kind of that backdrop and there's that ruling to make it so that they can work together well and so that it isn't every single... There's kind of the experience plain aspect as well of Data Mesh of that each time you go to a data product, it's not a completely unique experience. And there are things where you're saying, okay, how is data quality actually measured? This data product might say, Okay, we're focusing our SLAs on X and Y and Z. But here is how X and Y and Z are measured in general across everything, unless we very specifically call out, hey, our point isn't that this is super accurate versus it is super consistent. Right? So you get the directional measurement, and so we understand that, we're like, okay, it's kind of like latencies. When people talk about p99 latency, the way you measure that, it's actually weird because every 99 or every 100 measurements, you're saying, what was the 99th, what was the highest one? And then you start to average those, and so p99 doesn't actually mean what p99 means for most people. So you get into that specifics. And I'm going on way too long about this, but I think it's important to have

people understand that it isn't kinda chaos with governance, it's that there is a set of policies and best practices, and that you make it so that it's easy for people to actually leverage the governance. You make it so, okay, I can just click one of these columns or PII. Adidas actually talked about they have it kind of backwards, where all of their columns when they come on are marked as PII, so you have to unmark them. That way nothing goes through that's not PII, right? So you might say, oh this column didn't get unmarked, but it's kind of an interesting approach.

So again, I'm going on way, way too long, but I think this is an aspect that people haven't really dug into to really talk about the specifics around the language, which can get a little frustrating, but I think it's important to lay that as the backdrop. So I would love to jump into the conversation around maturing somebody's governance, right? Not getting any specific to any one client or anything like that, but you talked about the pathway is a lot of times you come in with no governance. Can you jump straight to this federated computational governance, or do you have to start to centralize so people can understand how governance can work and why the centralization is actually a pain point before you can move to federated. Can you start to talk about what you've done with Data Mesh and what's been successful, and what are some maybe antipatterns as well to avoid. So very, very broad question there that you could probably all talk about for 20 minutes, but let's start to head in that direction around the conversation.

0:21:51 Shawn Kyzer
From my perspective, when I think about the greenfield versus the brownfield, in a way, I actually find it easier when we come into an organization and they don't have any at all. Because that means we can kind of say, 'Okay, well, this is what governance is." We will have multiple sessions where we'll kind of train. We have different things that we'll do and exercises we'll go through with them, like various governance activities to be thinking about what the types of things they are going to govern and what is just enough governance in the federated model and how that will affect the various data products, right? And so I think for me, it's much easier to come into an organization that doesn't have much governance and then help train them up on governance.

Now, the downside to this is, there's quite a lot of work to do, right? So you need to, first of all, explain to them that they need a governance team. And one of the mistakes, not mistakes but one of the things that I commonly see that holds it back, is that maybe people are 10% or 20% dedicated to the governance team, and the reality is there really needs to be one or two people who are 100% dedicated to forming the governance policies and running the different sessions that you have, and just kind of building this documentation and touching base with everyone. Because governance is first and foremost, basically you wanna keep the company

out of trouble. So if you think of it from that aspect, and you can kind of start there with them and just be like, "Okay, what are the basic things that we need to do to secure the data? What are the basic things we need to do so that everybody can trust the data and the metrics that come out of that," are things that we can use to make meaningful business decisions. So then you bring in data quality and metadata management, that kind of thing. And then also all of the framework of Data Mesh can drive different policies within the governance board or within the governance team. What do you think, Gustavo?

0:24:10 Gustavo Drachenberg
Well, I think something that Scott said made me think that, yes, the term governance sounds boring and old and big, and when they do it in practice, it's actually, it's fun. In a sense, it's not... Well, it is hard then to fit all the moving pieces, but in a sense, it's quite practical, it's not as high level theoretical as we might think. When you think of governance it's a big word. But it's actually, it's fun. And as Shawn mentioned, we started with an MVP of governance sport, so what would be the MVP of a governance team for this company. And from there, we help them just as Shawn mentioned, to come up based on a list of categories that we know that are common in data governance, which are the ones that they consider are where the top needs are in their organization. And from there, we created a backlog of just policies that we needed to create and draft. And at the beginning, Shawn and I were leading this board with a client, and then we set up something very simple in terms of a trailer board of just policies, like a backlog, and then the ones that we were drafting, the ones that were in progress and the ones that we're reviewing. And then once that was done communicated and then stored in a common place and location. So the way it worked is very common to this board, which is a team operated, it's just like any other team would, but they were just creating policies.

0:26:01 Shawn Kyzer
Yeah, and I would like to just kind of highlight that the communication piece is really important because once they developed a policy, it's really important that there be solid communication to everyone in the organization that this is the policy. 'Cause I've seen a lot of situations where maybe a policy gets created, but it's not well communicated, then no one knows about it and so no one follows it. Right. So that comms piece is really important, the interface. And some of that can be done at the team level, where you have champions or local representatives that are embedded inside the Data Mesh teams, and they interface with the governance team. And so even though they're decentralized, there's still this really a bit of a central or federated governance team that they're interfacing with directly, so there's a connection there.

0:27:03 Gustavo Drachenberg
When we talk about a team, this is sort of a cross functional team, let's say, having

roles from different areas. So we have representatives from product, so we had a head of product, and if they were too busy and they couldn't join because this team met biweekly, they could appoint a representative. But at least there was somebody from product voicing the needs of product or the view of product, same for legal, same for security, same for platform, same for just the data organization. So when we had to draft a policy, they would be the champions of that policy, but they would take it back to their teams, get feedback. The other parties would contribute, so it's not that just this team in isolation was setting down the policy. They were leading the policy creation, but creating the policy was collaborative with the main representatives of each area.

0:28:06 Scott Hirleman
One thing, so Laura Madsen had talked a little about this, that the governance committee is typically very ineffective because you have a bunch of decision makers who don't have the context around this. And I think that that feels right when people really think about this. So how do you prevent the governance board from being that committee? Where there's either a thing of context exchange or there's doing work, but when you try and have high context exchange and high decision movement where, around people who don't have expertise around a specific thing, either you have somebody who's making decisions where they have no real idea of the impact of the decisions, or you have people that are very, very hesitant, rightfully so, to make these decisions, 'cause if these are big impact decisions, you're heading people for a potentially bad path.

So would love to hear how you've done that and then as well, we talked a little bit about the greenfield, but would love to talk about how you're also... Maybe you can weave that into the brownfield. 'Cause when you come in and people have these committees, they're already non functional, they've already been coming up the work so bad, and that's why people, when they think of governance, many people literally physically shutter, If somebody is used to trying to do the work, and then they start to talk about governance, ineffectual governance creates more harm than good, even though it does do risk minimization but it doesn't add any value. It's only about preventing cost or preventing risk, it's not that value add and we want to head towards that value add.

0:30:05 Shawn Kyzer
Yeah, I think, yeah. So once upon a time, I was very much a part of setting up a traditional data governance board, right? And we did do a lot of that, it was all about risk mitigation, that kind of thing, and in the end, it started to feel like, I would say like a rubber stamp organization, like things would just come through and you'd be like, "Okay, okay, okay." Yeah, it didn't give that value add. And I would say that one of the tricky things, this is why I said greenfield is a lot easier, is because when you go into

kind of the brownfield where this kind of organization already exists, you almost have to decompose it in a way, or you don't wanna completely dismantle it obviously, but you need to change the way it works, and the first thing is ensuring that you have the correct representation. So just like we have representation in our government bodies, you also need representation in your federated governance board. Right. And so you need those local representatives, but you also need people who understand all of the dimensions of data governance, and I think that's what's frustrating when we talk about governance, because if I'm a security person, I think about governance from a security perspective, right? I don't think about it from a metadata management perspective or from a platform perspective necessarily.

So you need each of these kinds of people to be represented in a diverse way on the team itself. Like legal and GDPR, even though they don't do anything technical, they absolutely have to be represented in some way. Just as an example. And so the tricky thing is you really need a top down mandate when you have to reshape, right? You need upper management support, when you have to reshape an already existing committee or board, whatever you walk into. Otherwise, it's very difficult because these people are already decision makers at the highest level, and so you might encounter some political pushback that can be pretty brutal.

0:32:23 Scott Hirleman
Yeah, I'm gonna say might is probably an understatement. And I was writing down just the thought of, you've got scar tissue to break through, you've got unlearning, right? It's easier to learn things right the first way than it is to have to change your habits, especially when the governance board often wields power. But if you've got the central governance team that's making these decisions without the proper context exchange, it doesn't work. So Gustavo, I would love to hear your same thoughts and what you've seen, especially around that brownfield.

0:33:03 Gustavo Drachenberg
So I think one of the pieces that might be missing very often, and I think it's the computational part, that federated computational brings in. You have this board creating policies, but then it's like you should comply with this, the whole team. And the team is like how? They're very busy, they're overloaded and it also adds a lot of pressure to the teams. But when you add the computational part from the platform already providing the teams the tools that they need, they're like, "Oh, if I wanna comply with it, here are the tools, I just have to integrate it with my data product and I'm fine." And it's a benefit for the organization and for the board as well, that scout their policies being actually executed. And that adoption can also be monitored, you can potentially automate and have reports where you see how much your policies are being followed in the teams, since they're using the platform tools as well.

0:34:09 Scott Hirleman
Yeah, I talk a lot. And something that's come up that I hadn't really thought of when I started doing this podcast was how crucial reuse is, especially because what you just talked about, Gustavo, is the concept of reinventing the wheel, right. How do I comply with this policy? I don't know, versus, Hey, here's this thing that it shows you exactly what is. When I talk to people about what is the data product? And I am super frustrated, everybody listening, publish your damn actual internal definition of what you call a data product, and there's a difference between the technical manifestation and what you would go to for a data product owner for the actual concept of what is a data product. Because everybody has to invent this stuff from their own head, and so then one, it doesn't look the same as you go from A to B to C, because everybody has a different interpretation versus if you have that platform capability, but like you said, you just plug it in and you say, Are we meeting our data observability goals, are we meeting our data quality goals in these different aspects? What are our SLAs?

And then it's like, we're measuring if we're hitting our SLAs. When Emily Gorcenski was on, we were talking about the error budget around SLAs and SLOs and things like that. And it's like, are you compliant with this? And if you're not, like, let's get compliant. Dave Colls was on. Yes, I've had way, I shouldn't say way too many, 'cause I love having the Thoughtworks people on, but I've had a lot of Thoughtworks people on and there are more to come. But Dave Colls was talking about fitness functions and that you can start to have that as part of your compliance or your governance of how well are we actually complying with policy, and that that can go back up to the board and say, "Do we need more movement?" But one thing that, again, Shawn, you kind of shyly touched on how political this stuff can get, but when you've got this board that's been running, they've had the power, they are the decision makers, but so often they haven't actually been making decisions or moving things forward because people have lacked content.

So I would like to dig into... When Shawn and I were talking about doing this episode a while back, we talked about how do you set up your data governance and then actually be successful? So I would love to hear about how do you keep the politics out of it, or how do you keep things moving forward so that you can actually make policy decisions, and then something happens and it doesn't grind everything to a halt. How do you keep the politics out of it? And I know it's different for every organization and all that, but people are frustrated by this. And then after you answer that, if we can set up ourselves to answer the question of how do you get going. Do you have to fully bake all of your governance policies that you'll ever have at the start, obviously we know that's not the answer, but how do you get comfortable enough to move forward? So let's start with combating the natural gravity of the governance board to become a thing to gum up the works, and that you can make those policy

decisions and keep it moving forward.

0:37:45 Shawn Kyzer
Yeah, I think, one of the aspects I like to focus on is high value policies that positively influence the lives of the technical folks, like the data product developers, and I'll give you an example. One is, for example, interoperability. So we have all of these data products and they all need to be interoperable in some way. Right? And so if we don't write a policy at the governance level that talks about what this interoperability will look like, then the data products are not interoperable with each other, right? So maybe they all must use a standard interface of some sort, right? Or APIs must have these following things and it must be available backstage or all data product output data ports must exist in a centralized data catalog. Now, these kinds of policies are perfect in the federated space because you can automate them, you can put them into tools, and then you're also saying very abstractly, you're not saying use tool X, tool Y, tool Z, you're saying it just must be a centralized data catalog. It doesn't matter what you use. Linked to Data Hub or Collibra, let the developers decide that.

Once people start to realize that there's this feedback loop between what the policies are doing and how it's positively affecting their lives on a day to day, if you start with those kinds of policies, suddenly the organization is like, "Oh okay, we understand why we need governance." And so if you can start with these high value items, and among them also security. I think security is one that may not necessarily make the data product developers lives too much easier, but it actually does help guard the company from different things that can happen with audits and different things with GDPR, that kind of stuff. And so that can actually influence other stakeholders, but you just need to say for every policy that you create, it needs to, in some way, add value. And you can almost quantify things like preventing the company from getting fines from GDPR or whatever, right? You can look at it from that respect and just think like for all of these policies, instead of just thinking about mitigation, think about what value it adds to whom. So it's not necessarily just security, but also different things like the consumers of the data, they'll discover that when they see the data quality, and that is a policy that it comes down from the board, when they see that, and whatever tool you choose, they'll be like, "Wow, I really appreciate having that because now I know I can trust the data."

0:40:56 Gustavo Drachenberg
Also these people are very busy. The ones that are part of the governance board, they are top people, so I think it's in their best interest to delegate a little bit of the policy creation. So the way we were doing the policies, I think was, I'm not gonna say quite fun, but it worked very well in the sense that the board would prioritize the things that they would consider important in terms of the business strategy and also the needs. And from there, they would appoint reviewers and approvers from each of

these areas. So we would draft a policy, somebody will be the champion of that policy, and they would get feedback from all of the other areas. So for instance, from platform, from security and so on, and the governance board was more in charge of making sure that this policy was moving through its completion and its reviewing, and then at the end, they will be the final approvers of the policy. So that way you ensure that it's not something that it's so out there that it's connected with reality. But that it had the input in the buy in of the people that put it together and that later will have to comply with.

0:42:18 Shawn Kyzer
Yeah, I agree. Make them feel like they're part of the process of creating the policies as well. Right, so you're basically getting their buy in. Yeah.

0:42:30 Scott Hirleman
Yeah, and I like that concept because it does keep it moving forward, if you can keep the ego out of it, of somebody saying, I have to be involved in all aspects of this, but if they can say, "Hey, I'm going to appoint somebody from my organization, my team, to make sure that this is moving forward appropriately." And that they don't... And that they can talk to that person and get that context exchange around, where are the things that we really felt we needed to have in here. I like that of break it down, but still don't try and take the power out of the hands of the people that are up because that's going to cause that kind of feedback of people starting to say, No, no, no, you can't take anything out of my own hands. It's like, if you can incentivize them and make it easy for them to hand that over and that it's the best solution as well then that's good. It's always a little bit of a political minefield when you're first starting.

So let's talk about, Shawn, you brought up inoperability, this is one that comes up all the time. When we are thinking about how we can get to a place where we're comfortable moving forward, I've talked to multiple people on this podcast, I just had someone very recently, Martina Ivaničová from Kiwi.com, and she was talking about, can we really call what we're doing Data Mesh because of X, Y, and Z? And to me, they're headed and they're on the journey, they're headed in the right direction. When is the Data Mesh, a Data Mesh? When is a man, a man. How many roads must a man walk down before he can be called a man. Like it's these, does that matter, does that actually, I don't give a crap. People think I really do about labeling something Data Mesh versus not Data Mesh, and it's like, it's about the approach. What are you actually trying to do and are you trying to accomplish that versus, oh yeah. We're just doing the data lakehouse and just calling it Data Mesh. Oh, that's not a Data Mesh, but how far somebody is down that. It doesn't matter to me. But when people are thinking about doing their minimum viable proof of concept or their proof of value or minimum viable mesh, which can all intertwine. What do you think

about when somebody says, What do I need to do so I can actually move forward? It's kind of like what you were talking about with the policies. Start with the policy and maybe iterate. Can you give people the permission. I'm trying to do this with a lot of episodes, give people the permission to say that you're good enough and you can move forward. Like CYA is basically what I say, and then beyond that, you can figure out how to add value as you move on.

0:45:28 Shawn Kyzer
Right. So we mentioned earlier a policy that talks about what is a data product, right. And we kind of discussed that. Now, we have, interestingly enough, even though we have the definition in the book of what a data product is, even still, we have kind of typically put together a policy of what it means to be a data product. Right. And we do use the DATS like discoverable, addressable, trustworthy, secure, so on and so forth, right. Those are almost like our index. And then you kind of fill in, in more granular detail as it relates to the organization around those affordances. And that was something that we communicated out to the larger organizations so that they know that the thing that they're building is in fact a data product that lives inside the Data Mesh, right. And I think you could even go so far as to talk about interoperability within a document like that. What do you think, Gustavo?

0:46:30 Gustavo Drachenberg
I think it's really important. That was one of the first things that came up. Because Data Mesh, it's still new and people will be, well, I have this pipeline, this is a data product, and I have this other thing, it's a data product. And it's like, No, no, we need to standardize what it's understood as a data product in the context of Data Mesh for everyone. And it was one of the first policies that we created, and one thing that we would say is the governance board would... It's like if you wanna apply Data Mesh in this organization, these are the rules that you have to comply with, that's the frame that we set in terms of policy making. So we started, Hey, if you wanna do Data Mesh, and you wanna call something a data product, it has to comply with all of these characteristics. Does it have metrics, is it observable, is it discoverable, does it consume, all these things that we know that a data product has. And from then on, keep on adding more context.

0:47:34 Scott Hirleman
Did you have to do that before you started to create initial data products or the chicken and egg of like, What do I have to have in place on my governance side. Because I think the CYA of having security rules. I even tell some people if you're really, really struggling with security, literally say your initial data products can't contain PII at all. There's no access rules, there's no nothing, because you're just figuring it out. And as people say, I need this PII in there because it will add value, then you have very specific ways for them to get access, but you can't get yourself

into trouble in general. If you don't have any PII in your freaking initial data products and it's as you learn how to share. And that's a terrible policy for certain organizations, 'cause the PII is the thing that really matters. And by the way, hopefully, we've said it a whole lot or I've said it a lot, but PII is Personally Identifiable Information. I assume most people know that, but especially if you're listening to a governance specific episode. But what is okay to get started? Where is it, okay, what's the line of how mature somebody has to be to actually get going?

0:48:55 Shawn Kyzer
Yeah, I was just gonna say, when I think about how we started, how we kicked these off, we concurrently started to build a data product by the way, with no PII also. That is a common pattern. I just realized that 'cause I was thinking back on multiple projects and I was like, "Do we ever start one with PII?" And we haven't. Because you wanna do something really simple at first, right? You don't want to do something hypercomplex or you need encryption, and you need to do all sorts of things with this PII, right. But as we were building the first data product in what we call the incubation period, which is the thin slice where you build the platform and the initial data product and the orchestration, all of that. We also kicked off the data governance board at the same time. And everybody was working very closely together for that first initial incubation phase, we learned as we were also building the data product, what kind of policies we needed to put into place. And because we had that, we had people that were working within the data product team, but they were also on the governance board and kind of vice versa and people on the platform that were involved in the governance board as we were building this data product for the first time, we actually just kind of learned along the way, it just seemed to happen organically what policies made sense. There were the given ones like security. Okay, right, cool. We got that right there. Especially there's experts that do that and we understood that.

And we also knew we needed to kick off a thread to ensure that we were doing GDPR correctly, right? But that happened a bit separately from the other policies like data quality, metadata management. Those policies just happened organically as we were building the first data product in the incubation period. And I think that's one reason like the incubation period can take so long, it can take like maybe six months to build your first data product, because you're doing the thin slice where you're setting up your initial version 1.0 of your governance work, version 1.0 of your data product, version 1.0 of your data platform. And yeah, it just flowed organically by making sure that all of those people were involved in some way in the governance and in the initial building of the data product and the platform.

0:51:30 Gustavo Drachenberg
And that came out, out of advice that Zhamak had given very early on, of one of the

things that was going wrong with implementing Data Mesh, which was scaling too fast. So she was saying, start with something very small, as small as you can, and then pick it from there, and that was such good advice. So we said, okay, that sounds good. And what was the MVP for the platform? What's the MVP for the data product? What's the MVP for the governance board? And that was our first insight. And as we were designing let's say what the MVP would look like for each of those, we were cutting back, taking it to the smallest and simplest things that we can do. Even that was very hard, to be honest at the beginning, very hard because it requires so much coordination between the team that has to build something, but it actually needs a platform to provide those tools and those things that we need to be agreed and coordinated. And development has to go, Wow, maybe the environments are being developed and so on.

And so things are happening, again, the governance board is starting and is meeting, you already see what is going on. And the cool thing about having a small team that's working closely together, even with governance, is that as platform people were involved in setting the policies, they will already be... It's like, Oh, this is coming up, I can already start thinking of what tools can I provide to enable people to comply with the thing, because they were already in the conversation. It's not like the data governance board showed up, and is like, "Okay, give me tools to comply with this." They were involved in the conversations from the beginning and they were already thinking, "Oh, we can do this with that, that, and that." So when the policy was out there, they already had an idea and a roadmap of what tools that we're gonna roll out to enable people to comply with those policies.

0:53:30 Scott Hirleman
Yeah, and I think that the time frame might have been a bit of a hidden lead for a lot of people, because I think we've had people on the podcast who they've done their MVP in six to ten weeks. And I think what they are typically proving out is slightly different, and I think it's valid to do this. I think it can set you down a path to a little bit more trouble, but what they're doing with their proof of value or the proof of concept. If you're proving out a data set is valuable, you're headed down a bad path. To me, that's just a terrible antipattern, but some people are proving out that they can create a data product rather than they can create a Data Mesh. Minimum viable mesh versus a minimum viable data product, and do we have the capability to even create a data product even if we are creating things that are kind of purpose built, but hopefully reusable on the platform side to support this singular data product versus minimum viable mesh.

And I think it's interesting, I haven't heard that Zhamak has said that. I don't know if that was in private channels or anything of people trying to scale too soon, because we've seen people that have been successful. A lot of times they had... Khanh Chau

comes to mind at Northern Trust. He talked about how they were doing a thing with data services, and it was taking them two to three months for each new data service, and so they were relatively mature in the concept of creating what should be data products, but weren't. Because then the consumer still had a very high amount of work and total cost of ownership to actually take something from the service and actually consume it in the way that mattered, and then their data cleansing and quality and all that stuff wasn't getting pushed back into the service, so if another person consumed from the service, they had to do that same work, which is again, what we're trying to not do with Data Mesh. But they had already matured to a point where they said, "We need this, they've thought through a lot of the things." And so when they moved to doing their POC, it was a much shorter time to become mature. They already had a lot of the conceptual infrastructure and they'd done a lot of the work, it just hadn't been paying off, and then once they moved to Data Mesh, they were like, "Oh crap, it's not two to three months, it's two to three weeks." And it's like the quality of these is 5X. The total cost of ownership for the consumers of what the data should be to when the point of consumption is now zero, instead of they had to extract the data and then prepare it for their own consumption. That you can push that more and more upstream.

But I do think that there are people that are really looking to jump too soon, and again, sorry I'm talking too much 'cause you're just sparking way too many. And this ties together, like I said, this is probably close to interview #90, so I've done a whole lot of episodes and it's tying together so much because so much of what you're saying makes so much sense of things that are just falling into place. So I would love to kinda talk about how you have this tight coordination, there are a lot of organizations that aren't mature enough to do that. So if they were to say that they wanna head down the Data Mesh path, what would you say to them? Would you say that you have to mature your organization to a place where you could have this tight coordination for six months, or would you say figure out how to do your data products and that you're gonna have a little bit on the platform side, but it's really... Look for the reuse, but it's okay to not be that mature at that point, what would you tell them? And not saying that this is the canonical Thoughtworks answer or this is the only answer for all time, but what would you push back on them if you saw somebody who wasn't mature enough to have that level of coordination for that period of time, 'cause most organizations I don't think can, they'd devolve into chaos.

0:57:57 Shawn Kyzer
Yeah, right. Because the argument is always, "Well, we have business as usual that needs to keep... The show must go on, right?" I would say start small. Try to pick one to two people from each of the layers, right? So if it's a house at the top, you have the governance, the roof protects the house, right? In the middle you have your data product developers, so they're living in those different rooms, which are different

data products. And at the base you have your foundation, so a couple of platform developers. And, at the very least say, "Hey, push for this small team just to bootstrap or kickstart your Data Mesh." And so maybe you can try to negotiate just a smaller group over a period of I would say... What do you think, Gustavo? Like maybe a quarter you could probably get away with three to four months, something like that.

And if you pick a simple source aligned data product that doesn't have PII, that would be a really great starting point, that's a way to get your feet wet, understand every layer and then also bring people from all those separate layers together, so that they understand and then, when they finish this, they go out and evangelize it to their respective teams. So you're training the trainers in some ways. And so, in that way, it can spider through the organization. Like if you just wanted a way to kick start something and bring in. I would also say that we do a lot of internal evangelization as we go through this. I've done many different kinds of presentations of, "This is where we're at, this is something we did on governance," or, "This is how we synchronize the platform and product teams." And we always talk about it, we put it in the Slack channel, we do different kinds of videos or whatever so that people can watch and they know what's going on.

1:00:21 Gustavo Drachenberg
Yeah. I really like when you said, Scott about the different levels of POC, an MVP of a Data Mesh, 'cause that's the way we've also seen it in practice. There are some clients and maybe they're not sure they wanna get a taste of Data Mesh, so we do a short POC which might involve doing a data product with some basic capabilities, which is what Shawn mentioned. And then, you can do an MVP for a use case that might involve several data products. That's another level. And then the other one is an MVP of a thin slice of what will be the Data Mesh. You will do an MVP for use case, an MVP of governance, and the same for a platform already performing as a Data Mesh platform team.

And also to touch on what Shawn said, and I was thinking as you asked that question about where to start. One of the challenges that we find as well, if you wanna do microservices today, everybody knows what a microservice is. They know the APIs or different technologies, everybody is well aware and it's just a conversation of, "Alright, how do we decompose this and how do we implement it in our organization?" In this case this is so new that we spend a lot of our time explaining teaching, coaching, and then you have maybe a little piece of the organization that they have the right mindset and so on, and then how do you disseminate that knowledge throughout the organization? So, currently is growing organically within the Data Mesh communities and also within an organization that's implemented Data Mesh, we try to pair people from another team so they can start learning on how they do it, so then when they move to their domain, they have the knowledge.

I personally had to coach several teams because we have product managers that need to learn how to manage data products using product management, then platform. Shawn worked very closely with the platform folks and data product teams to help me figure out the right architecture. So it's really a combination of educating, teaching about Data Mesh and the concepts, and then now that you understand what it is, how do we put it in practice within this organization, and at what level? To do a full scale implementation you really need a big commitment from leadership because this will require an investment of people that will have to be allocated to that. And again, if you were running a monolith today and you wanted to decompose those into microservices, you would have to add additional capacity to do that work. It's not gonna happen with the same people that are keeping it running.

1:03:04 Scott Hirleman
Yeah, I talk about if you're giving people additional responsibilities without additional resources, that's a... 'not nice' move. I'd probably use a different phrase than 'not nice' if I weren't on the podcast. And it's interesting because I would look to the pragmatism in practice episode as well as my episode with Scott Hawkins from ITV, but, I can't remember his name, but the Head of Data at ITV was on an episode of ' Pragmatism in Practice' with Kimberly Boyd and Danilo Sato. And, ITV uses the consultant in a box model, or a consulting team in a box model. So they actually have this group of floating, 15, 20, 30 people that they bring into a team, and so it gets them to producing a data product sooner because they're the ones who are doing a lot of the development and maturing that team very quickly. And so they've got this kind of capability that they can drop in a lot of places, and it doesn't sound like you're doing that quite as much as training. This is somebody who's coming in and doing the actual very specific work and is very focused on upskilling and then just moving team to team to team,

And then as well, Shawn, you've kinda tipped yourself against... Once I finally get my survey out there, which I've been working on this for too long and I need to just get going. But my 'Getting Started Survey', you've tipped your hand on one of my key questions of how many data products do you start with? You've said one over and over. There are a lot of people who say, "No, you've gotta test it in interoperability or you've gotta find a use case." And there's this emerging pattern that I think has some problems but of creating a purposeful data swamp where you share your data so people can see what data could be on offer. Use cases emerge from that, and then, they push it back into the data producers and say, "You need to serve this use case," but that's a consumer driven or a consumer aligned data product when there isn't a source aligned data product to it.

So that can cause issues, 'cause you have to create micro source aligned data

products where you might only have a little bit of data that flows into this thing and you've got multiple of those and it can cause issues, but it also gets you to a high value use case where you know there is a direct registered consumer upfront. So it's a lot of interesting things again, where you're tying through a lot of these really helpful patterns to think about. And I think a lot of what you're talking about... And, when you talk about the big commitment as well, something that's come up for me lately is Data Mesh, a lot of this is, people think about it as an initial investment. The initial investment is to build momentum, and you have to start securing wins so that you can prove your value to get incremental investment 'cause it's continuous investment. So you need to drive continuous value, but a lot of people try to go for too big bang of an approach, so you have to ask yourself internally, "Are we going to be able to stick this out for six months with a decent amount of investment?" If you're gonna be proving out minimum viable mesh from the start, versus, "Can we get to just a smaller data product that's gonna prove this out, that can prove out we can do this?"

So, again, you're just tying through a lot of things. I shouldn't be taking up all the time on this, I would actually like to do a followup where I just do a 15minute summary of this episode or whatever. So, we've talked about a whole lot of different things. I think a good place maybe to wrap on, unless you've got other topics that you wanted to make sure we cover. But about, what part of governance do you think should live in the platform itself versus at the product level? It sounds like you're saying the decisioning, the policies should be at the platform and they should be enforced via the platform and offered as affordances via the platform, not just, you comply with this, but did you comply check box, but that you make it so that people can comply easily. But, what do you think lives at the product level versus the platform level?

1:07:41 Shawn Kyzer
I'm gonna be opinionated on this and say, as much of that as you can meaningfully perform at the platform level I think is the better approach. That being said, there are certain things like, think about tagging of the data, sometimes that's something that someone at the data product level, they know the data, they're part of the domain, only they probably know how to appropriately tag the data. That's probably not... You can make it available that you can add tags at the platform level, but at the end of the day, the data product team is going to have to fill in that.

So, anything that's specific to the data or nuanced, it's probably gonna be at the data product level, but those things that are more general such as security, encryption, that kind of thing, those are gonna definitely gonna be at the platform level. The reason I say this is because there is this... So, you're the governance here, right? So not only are you responsible for creating policies, you're also responsible for

monitoring that these policies are being followed and for a certain amount of auditing. So you need visibility, and it's much easier to get that visibility at the platform level than it would be to try to extract it from each of the individual data products. So think about access, control of data sets and different logs like that. Most of that, they're gonna need to prepare reports or have dashboards available from the platform up to the governance level.

1:09:21 Gustavo Drachenberg
So, suppose... Again, coming back to the larger schema, the governance board said, "You need to have data quality measures. Some SLOs." And the platform will provide these tools, but then suppose you're building a data product related to financial data, completeness needs to be 100% because you need to account for every penny. But maybe, if you're building a data product that you just want to provide information for a product manager to make a decision in just an overall trend that is happening with the application very quickly in the final or something like that, then freshness might be more important for that data product than completeness. So that's where the product teams get to customize what makes more sense for them.

1:10:20 Scott Hirleman
Yeah, I think that determination of SLAs, Emily Gorcenski in a webinar from mid last year talked about, they had one place where they had two different data products that were the exact same data product, but one had very very very quick freshness. I don't know how you say high versus low freshness, but it was like a five minute freshness. But it had low accuracy, 90% accuracy 'cause people needed it with that. But then there was another one that was 99.99% accuracy but two hours freshness. And, instead of trying to get one to hit both, that would have meant 3x the cost from having two that are doing the same thing. And so, I think that's very crucial. So again, I do apologize to listeners, I wish I would have spent more time asking you questions, but so much of this stuff is just... You've dropped so many things into place for me, so I really appreciate that. Is there anything that we didn't cover that you think we should have? Whether we could do it on a future episode or anything that you would wanna do that, or is there any way that you'd like to wrap up the episode? Any kind of button or anything you'd like to put on the episode?

1:11:39 Gustavo Drachenberg
So, I would just like to mention one thing that we didn't cover, but I think it would be interesting for the listeners is team size when you're thinking about developing this MVP for a mesh. It was having three or four people in the governance board, a team of maybe six data engineers and our product manager for the data product team and maybe four people in platform with strong leadership platform roles, that could be for six months a good team to create one or two data products to have an MVP. But then when you develop the platform capabilities and we start to scale, the time

got reduced to maybe the first time, to three weeks and which is somebody that will be performing as a data product manager and maybe two data engineers, BI engineers. And then, as more capabilities kept ongoing like the times got even shorter to produce a data product. So, the amount of time at the beginning is a little bit painful but then you start to reap the benefits moving forward.

1:12:49 Shawn Kyzer
Yeah, I actually have something I'd like to piggyback off of. Yeah, that's exactly also how it works at the governance level. At first it really does take a lot of effort just getting everyone together, setting the ceremonies for the meetings, compiling the team, and thinking about how everyone's gonna interface and what the communications are. And building that first policy is probably the most difficult, and so once you realize how you want to operate at the federated governance level, and you start to understand the nuances between federated and what we think of traditional governance, then things really start getting to rock and roll. You start to come up with all these different policies, you start to get feedback from the platform and you get feedback from the data product developers about the kinds of policies they would like to see, and before you know it, you have a nice little backlog and you're just churning through things. But, it's not easy, and it's multi-dimensional. So, don't get discouraged, just head down and focus on the things that are going to yield the highest value at the governance level and help you enable your Mesh.

1:14:17 Scott Hirleman
And, I think one thing that's come through in a lot of conversations, including this, that could be set out loud is, it's okay to get things not right the first time, because the cost of evolution. I just did an episode about, "What have I learned?" The cost of change in data has been massive historically. Changing anything has been such a massive pain, friction, everything like that. And we have to get ourselves to a place where that isn't the case. But we've got scar tissue as well to move through. So that's when you're talking about the greenfield versus brownfield; their scar tissue. People have had bad governance situations and so they're like, "Oh, this thing is terrible." So, I think that as well that, "Don't get discouraged. It's okay to not get it right. CYA, cover your butt but from a security and that standpoint." I think that's what I'm getting from you. I don't wanna put words in your mouth but that's what I'm getting as well as. It's like, it's okay to move forward with a little bit of uncertainty. You have to embrace ambiguity and change to move forward with data.

1:15:31 Gustavo Drachenberg
Yeah, Definitely, the first talk that we had with our teams as we started is like play, learn, there are a lot of things that you're not gonna know and probably not gonna get right, but it's part of this. 'Cause it's something quite new at the moment.

1:15:49 Scott Hirleman
It's again, like if you created a data warehouse, the cost of change of it, an enterprise data warehouse is so high that Data Mesh is about getting rid of that cost. And so, we have to figure out exactly how to do that, but people are still just like, "I don't know. I don't know." Well, again, as you can probably tell from how excited I am, this was a phenomenal episode. I really enjoyed the conversation, and like I said, you crystallized a lot of my thoughts. I know you said a couple of times, "Oh, you made me think about this or this." But this crystallized so many things for me. I'm sure there are gonna be a lot of people that wanna follow up with you. Where is the best place to do that? Where do you want people following up with you about?

1:16:32 Gustavo Drachenberg
LinkedIn in my case, I'm available there.

1:16:35 Shawn Kyzer
Yeah. For me I think LinkedIn is the easiest way. And hey, if you need help telling your Data Mesh story, Gustavo and I are available so just hit us up on LinkedIn.

1:16:51 Scott Hirleman
Okay. Well again, Gustavo, Shawn, this has been so great. So thank you so much for the time today, and thank you as well everyone for listening.

1:16:58 Shawn Kyzer
Alright, thank you.

1:17:01 Gustavo Drachenberg
Thank you.

1:17:01 Scott Hirleman
I'd again like to thank my guests today Gustavo Drachenberg and Shawn Kyzer of Thoughtworks. You can find links to their LinkedIn profiles in the show notes as per usual. Thank you. Thanks everyone for listening to another great guest on the Data Mesh Learning Podcast.

Thanks again to our sponsors, especially DataStax who actually pays for me fulltime to help out the Data Mesh community. If you're looking for a scalable, extremely cost efficient, multi data center, multi cloud database offering and or an easy to scale data streaming offering, check DataStax out, there's a link in the show notes. If you wanna get in touch with me, there's links in the show notes to go ahead and reach out. I would love to hear more about what you're doing with Data Mesh and how I can be helpful. So please do reach out and let me know as well as if you'd like to be a guest. Check out the show notes for more information. Thanks so much.