accenture

EP. 30: WHAT'S NEW ABOUT DATA VIRTUALIZATION WITH DATA MESH?

VIDEO TRANSCRIPT

Bryan Aller [00:00:00] But you'll quickly see here how data mesh you could move and shake almost any industry.

Teresa Tung [00:00:12] Hi. Welcome to another AI Leaders podcast. I'm Teresa Tung, and I'm Accenture's Cloud First Chief Technologist, where I look at the future of cloud. This session is on data mesh and specifically about what's new about data virtualization with data mesh. Data mesh is a shift away from centralized data towards decentralized and federated data management, and it's driving a new interest in data virtualization. And so, I'm thrilled to be joined by two experts who are going to share their experiences with using this new generation of data virtualization for data mesh. So first, let's welcome Justin.

Justin Borgman [00:00:45] Thank you, Theresa. I'm super excited to be here. I'm Justin Borgman, co-founder and CEO of Starburst. Starburst offers a query engine that ultimately allows you to analyze data where it lives, which fits very well with this topic.

Teresa Tung [00:00:58] Thanks, Justin, and let's welcome Bryan as well.

Bryan Aller [00:01:01] Hi, I'm Bryan Aller, I'm a director of platform engineering at Comcast. And I am responsible for a portfolio of enterprise big data platforms, which we're offering as a set of hybrid shared services. And that includes an enterprise data fabric using Starburst data as well as a host of self-service frameworks and tooling. We've implemented data virtualization today and we are on the road to building a data mesh.

Teresa Tung [00:01:28] Well thank you both for joining us and sounds like we have the right experts to talk about how data virtualization fits into data mesh. We're going to be going through some examples really showing how this generation of data virtualization scales and why it's needed for data mesh. So, let's get started. So, first question, any data mesh podcast has to start with a definition of data mesh. So, Bryan, as somebody who has begun this data mesh journey, what is it and why are you interested in it?

Bryan Aller [00:01:55] All right. So, we're excited about data mesh for a variety of reasons. You may not often think about a large telecommunications or media company being the first out in front or talking about some of these types of technology innovations. But you'll quickly see here how data mesh you could move and shake almost any industry. So, the rough working definition for us, we look at a data mesh as a collection of data platform components, these nodes that are providing interoperable services in the form of storage, computation, transformation and egress, and most

importantly, the culture to support it. That cultural piece is paramount. Unlike many previous technology trends, this particular disruptor is more than just a technical solution so to truly realize the value of this as a solution, we want to pursue a fundamental shift in the way that the enterprise thinks about data, products and consumers. This has the ability to unite a variety of independent business units under a common governance model without threatening their engineering sovereignty. Virtualization plays a large role so this allows us to separate, compute and storage in the user interface and allow our users to still have some control over what they do today.

Teresa Tung [00:03:22] And I think that's why anybody who's worked in enterprise data has really embraced and is interested about the data mesh topic. It's this concession to reality, right? We've been trying to centralize data, whether from organizational structure of a system structure for a long time, and it's not going to happen. I'm going to be the first to say as we work with more lines of business, new technology and new partners. And so, I think what you described at Comcast sounds very similar to how we're seeing other interest pick up. And so, with that decentralization, it requires, as you mentioned, maybe new ways of working and new tools. So, Justin maybe I'll have you add what do people need to know about data virtualization technology and the role it plays in data mesh?

Justin Borgman [00:04:09] Yeah, great question, Teresa. And I love the way that you put it as a concession to reality. I think that's exactly the world that we live in today. I think for years, for decades, we've been trying to centralize all of our data into one central enterprise data warehouse with the goal of being able to understand our business holistically. And I think there's now a movement to sort of say, you know, that is not a practical proposition, and if I want to enable agility and self-service analytics for my organization, I need to be able to access the data where it lives. Now to your point, data virtualization as a concept is not necessarily a new idea. It's been around for quite some time, but the biggest thing that has changed is really being able to execute queries through data virtualization and do it in a performant way and at scale. Because the earlier generations of data virtualization really lacked the ability to run performance queries and run them at scale, and that left kind of a bad taste in people's mouth. And the reason for that was that those earlier iterations really relied on the underlying database systems to do all the heavy lifting. They were really more of query routing technologies than they were query execution technologies. And what's changed in the last 10 to 15 years is really two things. Certainly, network bandwidth is continuing to improve, so network becomes less of a bottleneck, but also the emergence of MPP architectures, massively parallel processing architectures that allow really fast queries at scale where the joins across data sources are happening within those query engines itself. We'll talk more about how that works, but that's really the fundamental difference. People will say, hey, we tried this. It was slow. It didn't run at scale, and we gave up on data virtualization. I think now is the time to take a second look at it, because really the technologies underlying it have changed dramatically.

Teresa Tung [00:06:11] Maybe in addition, with data mesh, there's new requirements. I think we have always wanted to be able to query data as a single pane of glass wherever it resides, and you mentioned it's now...this new generation is more performant than before and does scale, right? And then the second half might be, I often use the analogy of what we're seeing with data mesh is like when rest APIs came out. So, we already had web services and we might have had certain things from managing web services in terms of metering and monitoring those web services, but when rest APIs happened it was very much a business change to say these web services are now business entities. And in the same way we're seeing with data mesh, your data sets and your models are now becoming data products and so maybe having virtualization in the same way we have an API And the challenge has been that, you

gateway, there's a new requirement to be able to control both, in this case, the consumption and the production of data products. So, I think the mesh is also driving a demand that maybe we haven't seen as much as before. Bryan, you're implementing Starburst in particular, but data virtualization as part of your journey, can you talk a little bit more about why you needed it and what did you use previously and how did Starburst help?

Bryan Aller [00:07:28] So our data journey at Comcast has been a long one. We have traditionally been very much in the forefront of adopting new technologies as they come out. And over the years have seen an evolution using everything from data marts, data warehouses, data lakes. We've separated our storage and compute. We have virtualized different solutions. We've containerized applications.

know, with each of these evolutions, we find ourselves with either expensive cloud migrations. We wind up with legacy solutions. We never fully migrate from one to the next, because there are valid reasons why these platforms are put into play. In the past, a lot of our development and making sure that these platforms could interoperate was done through the use of ETL or data pipeline to help centralize data. And as time went on, new formats made that more and more difficult so structured data to SQL. Going from block storage to object storage. And we're now in a situation where there's a long tail of legacy data and the cost of moving wholesale from one to the next might be quite expensive. So, looking at what we've learned from the big data space containerization and how that's helped with cloud migrations, we're now operating in a hybrid environment, and the exciting part about looking at data fabric and data mesh technologies is that it seeks to drive harmony. It allows us to tap into all of these different data stores, level the playing field and enable interoperability between these different solutions, even if the underlying data storage formats are different and sit in different places.

Teresa Tung [00:09:17] I think the privacy part that I was following really close was very interesting. You mentioned things around regulations and needing to keep data where it's at, but you also talked about best of breed. Is there a sort of legacy reason or privacy reason why you might start with data virtualization in addition to the best of breed reason?

Bryan Aller [00:09:37] Yeah. So beyond it simply being best of breed, there are valuable benefits to both privacy and security to looking at a solution like this. GDPR and the California privacy law, there's more of a need now than ever than there was in the past to know where the data lives and to be able to govern it in all of those locations and how people access it. So, when we look at virtualization technology like this, it creates another layer in which we now are able to control, the end user experience. We're able to enact data access policies to show that even if data lives in different places across the enterprise, it can be accessed in a uniform way. It can be secured in a uniform way. So that in itself is a huge win because in years prior, we were in a position where governance had to happen on each and every individual platform, and there are many different ways to apply governance, and now the standardized toolsets that a fabric or a mesh bring to the table are simplifying that experience. We're able to implement policies in one place.

Teresa Tung [00:10:54] I love what you just described. Data virtualization is a great way to bridge that so you don't need to be of a certain maturity level before you add data virtualization. What you're saying is add data virtualization right away, use that to unlock your data products from your existing data systems, and while that's happening, you have this common facade that makes it really easy so as you're upgrading underneath the systems and modernizing your users, they get the benefit of the performance, maybe the benefit of the cost, but they don't actually see a change in terms of the data products that they're working with. So, I thought that actually a very powerful way to get started. I think, Justin, you did explain already the value of data virtualization in this new generation. And in

fact, at Starburst, you don't call yourself the data virtualization vendor, maybe because of that bad taste. You call yourself a SQL based MPP query engine and query acceleration. And those terms mean that it's been engineered to scale right from the beginning. Can you share some more about Starburst's design and heritage for this engine?

Justin Borgman [00:12:04] Yeah, absolutely. So, as we touched on earlier, the drawbacks to the earlier generations of data virtualization were really performance and scale, and that's precisely at the heart of what we deliver at Starburst. Now, the core engine itself is actually an open-source project today known as Trino, but originally called Presto. My co-founders are the creators of Presto and Trino and Presto was born at Facebook, so it had to be born at scale. It had to run at literally the largest scale imaginable, hundreds of petabytes of data, thousands of concurrent queries. So often, you know, when we're first engaging with the customer, there might be some natural skepticism of does this work? Can this really run at scale? And, fortunately, we just have to point to the giants who have who've really pushed this to the limits - Lyft, Netflix, Airbnb, LinkedIn - the list goes on and on that are running this operationally at tremendous scale. And so ultimately, that foundation is based on a parallel architecture, an MPP architecture, massively parallel processing, taking advantage of a cluster of machines really of memory and compute. They're able to execute that query in memory at nearly infinite scale. Add to that the fact that we have connectors to a variety of different data sources, almost 50 different data sources today, and many of those connectors themselves are parallelized. So, one of the ones that Brian's very familiar with at Comcast, for example, can connect to a Teradata to enterprise data warehouse. That's a parallel connector. We're able to make faster reads from those types of systems again, in the spirit of performance and scale. So those are kind of the core differences between what we do and those earlier versions of data virtualization. But, you know, the spirit of the value that we provide is similar, which is ultimately that you can run a query from a single point of access and have all of your data at your fingertips.

Teresa Tung [00:14:04] One other way of looking at scale, right, it's certainly the performance, but sometimes it's also the user enablement. So, I'm going to maybe start with Bryan and then Justin maybe you can also chime in as to how Starburst actually add that on top of what open source does. But Bryan, maybe you go first. Can you tell us about the user enablement you've done at Comcast?

Bryan Aller [00:14:27] Yeah. So, the key to being successful with a solution like this is to really enable your user base. In the past moving from one solution to another often times end users would complain. They'd have to go through lengthy training. There'd be a learning curve. And the biggest thing here in adapting to a culture and a cultural mind shift is showing how folks can pull forward the tools and technologies and processes of the past but shift the mindset away from thinking about specific underlying technologies to thinking about re-shifting and focusing on the data, focusing on where the data is making a difference. And so, what we've done is we've begun hosting a variety of user group sessions internally to educate on a query fabric, to talk about mesh principles. We've partnered with Starburst and brought them in to do instructor led education and provide our end users with materials and do workshops to teach them new skills. In a lot of cases, our end users have simply been able to just change a connection string and use the tools that they use today, whether it's a reporting tool, a dashboard tool, a spreadsheet tool, and keep moving. And for them, the experiences are relatively seamless, and that's the best thing that we can do for end user enablement. But in cases where that's been more of a challenge, we've been able to document job aids, document different reference architectures or sample code that make it easier for them to adopt. And so far that adoption has lowered the barrier to entry to a point where folks that are less technical, even folks that are strictly in management and have

no technical background, are actually able to jump in and leverage data in the big data ecosystem.

Teresa Tung [00:16:33] Yeah, Justin, we didn't touch on data products and how Starburst supports that, but it's really in addition to the scale and the new types of architectures ,the connectors… you've really focused at Starburst a lot on data products and making it easy for users like Bryan and his team. Do you maybe want to touch on some of that?

Justin Borgman [00:16:51] Yeah, absolutely. I'm so glad Bryan mentioned that because we really think it's one of the absolute keys to having a successful data mesh strategy because it speaks to the people and process side of data mesh. You know, when Zhamak Dehghani first coined the term data mesh, she referred to it as a socio technical phenomenon. I had to go look up that word. What does socio technical mean? But the socio part is referring to the people aspect. There is a people and culture change that Bryan alluded to, and so how can you help facilitate that? Sometimes that's even harder than the technology side. Well, I think data products is a big part of the answer to it because data products basically kind of changes the nature of the conversation. It's less about where is it my data physically lives and the speeds and feeds of how I access it. It's really about how do I create data as a product that is valuable to the rest of the organization? How do I curate and manage a kind of gold star thumbs up quality approved product that others in your organization are going to want to consume? And it's no different than a lot of consumerization in technology, you know, broadly, which is it's a marketplace, effectively internal marketplace, perhaps someday external marketplace, but internal marketplace for the various groups and teams that Comcast and other organizations to be able to share the data that they know and that speaks to another core pillar of data mesh, which is really domain ownership. The people that know that data the best should be responsible for or involved with the creation and curation of high-quality data products to be consumed. And I think data products is a way of really making that super simple for folks and for people to see the value of all the work that you're doing under the covers. I can go into one interface, find the data products I'm looking for and start to consume them. And, at the end of the day, that's what this is all about. We're all trying to move in the direction of self service and data democratization and really putting the power of data in the hands of everyone from the intern to the CEO. I think data products is a great way to do that.

Teresa Tung [00:19:05] I mean, I think data products is what the business cares about. Maybe data mesh is one way that data geeks like us might be rallying around how to do it.

Justin Borgman [00:19:16] Yeah.

Teresa Tung [00:19:16] So I'm going to ask Bryan, like why? So why should we get started with data products? Why should people use data virtualization? You have a journey and you have scale at Comcast, but how did somebody get started? How did you have to convince?

Bryan Aller [00:19:32] Yes, there was definitely some convincing that needed to happen, and I would say that the story speaks for itself. It sounds almost like it's too good to be true. When you mentioned that you can bring in a solution that promotes value across everything that you've done to date and future proofs you and offers performance gains, it seems like a lot. But being able to bring that to the table with quantifiable examples definitely helps to sell the case. There's the obvious privacy and security benefits. There are huge performance gains, especially with big data, you know, 10 to 20 times faster on high on SQL queries is nothing to joke about. And we've demonstrated through proof of concept initially that we could take a best of breed approach and leverage our existing data stores and join them against new ones. And by putting those live demos in front of senior leadership, it was easy to make the case that this really should be the broader data strategy. You know, it's not starting over. It's not

a costly migration. We're able to avoid and defer those costly migrations, and we can prolong the value of our current solutions in conjunction with supporting a forward-facing solution that supports our competitive edge. As new technologies come along, we can bake them into this model. It sets us up to harmonize our overall ecosystem. And last but far from least, it sets up a shared context for identity authentication, authorization, data, access policies. So, you know, the benefits far outweigh the costs. It simplifies the ecosystem. It makes for a better customer experience, and it sets you up in a position where you can make change today. You're not starting over. You can take advantage of these improvements immediately.

Teresa Tung [00:21:34] There's a lot of reasons, it sounds like why to do it. But summarizing it's the data product mindset and the ability to really democratize the use. It's that easy button, right? Like you mentioned that you don't need to... before data virtualization, I would have had to gain access into each of those source systems, possibly talking to somebody, making a copy of the data or getting direct access to it. And I think now with this sort of product mindset and the only way to scale is really to distribute that out, right? And so, data virtualization means that they can still control the access that the data owners and the product owners really need to set as well as balancing with the self-service model, right? So, I think that why anytime you have different systems, which I would challenge any enterprise organization does, anytime you have cloud migration, cloud data modernization, maybe those are good places to think if you have these projects. That's when you should start. Justin, first anything to add, and then how do you start? If we know there's lots of reasons why you should start, how do people start?

Justin Borgman [00:22:50] Mm hmm. Well, I think the easiest way to start is just, first of all, keep it simple. And I think what we often see is people simply connecting to data sources. And very often it's a data lake of some kind, much the way that Bryan discussed object storage. I think every company should have a data lake somewhere just because it's basically free. It's almost like that storage that Apple gives you with your phone, you know? The free storage, right? It's the cheap storage. And so it's always a good idea to store as much data as you can in a data lake. And we think pretty much everybody has one, but then reaching beyond that data lake and connecting it to an additional data source is a great way to start to see the benefits of virtualization. So, in Bryan's case, it was a data lake and a data warehouse. For other customers it might be data lake and database, an operational system perhaps. It could be connecting to Elasticsearch, connecting to Kafka, connecting to another data lake. We have a lot of customers that have multiple data lakes. That's also a great use case. So, I would just start simple is kind of my main advice. Connect a couple of data sources, start to run queries and demonstrate the value of being able to get the insights you need without having to do any data movement in advance of that. And then from there, I would start to really think about data products, which we spent a lot of time talking about here, because I think that's a great way to connect the technology to the business value. I think, Theresa, you put it very well, data mesh might be for us technologists, you know, the geeks in the room, but the business cares about the value that they're getting from it ,and data products is a great way to make that connection, for US technologists out there to help sell your own bosses, to sell to the line of business on why this is a worthwhile investment and what it can do to really transform your organization. So those would be my two pieces of advice start small, connect a couple of data sources, and then start to create and curate some data products, and I think that becomes a great demonstration of the power of the technology.

Teresa Tung [00:24:56] Okay. Well thank you both. Thank you, Bryan, so much for sharing your journey at Comcast. I think the way that you've gotten started with, as Justin was saying, with the two systems and then moving to cloud and then going enterprise wide really does prove that this technology doesn't just work for digital natives but works at enterprise scale as we

know it. And thank you, Justin, so much for sharing, helping us understand what this new generation of data virtualization is and that it's real, it's performant. It was designed and birthed to be. So, thank you so much. Thank you, everybody.