A.12 Expert K

Interviewer
Okay, I started the recording. Yeah, to give you a short background about me myself. I'm Tom, 24 years old. I'm Dutch. I have a background in industrial engineering. That was my bachelor and I'm doing a master in data science and entrepreneurship. So it's about combining the business background together with the more technological driven background to make data driven solutions for business practices. At the moment, I'm really deep diving into data mesh and yeah, the goal is to create three different frameworks. The first one is about data as a product. The second one will be about the self-serve platform and the third one will be about the government layer we need on top of both to make sure that everything in the mesh is structured and interoperable, etc. Very brief background about me. Oh, and by the way, I'm interning at Bain, Bain & Company at the moment. The first part of my internship was in Amsterdam and now I'm doing the rest of it in Chicago. And what's the connection with AWS? I know somebody asked me to meet with you.

Expert k
Is it just friend type thing?

Interviewer
No, actually that was *. He is some kind of partner of Pyxis. So I'm working for Pyxis, which is a data analytics company and it has been acquired by Bain a couple of years ago. So now it's part of Bain. It's more like a department right now. We do a lot of things with AWS. Actually our whole data architecture is built on top of AWS. So yeah, indeed and * is our solution architect. So yeah, we're really fortunate to have him.

Expert k
So, how can I help?

Interviewer
Yeah, if you can perhaps share a short background about yourself, then I can show my frameworks afterwards and we can have nice discussion about it.

Expert k
Yeah, I So been in high-tech for a while, you know a couple firms, two startups and now *. Been with * for about five and a half years. Had a couple roles. Originally part of the strategic accounts team. Then I was running in 2019 I ran the AI ML business development team covering all of America's. Ran the worldwide * team for about a year and a half and I've been now with analytics for about a year and a half. My team runs worldwide go-to-market for 10 different analytics services. Services like EMR, Glue, MSK, KDA, KDF, KDS if those acronyms mean anything to you or initials. You know, I've got I also have Athena. I just recently picked up my team just picked up Data Zone as well and information in that flow. So, you know my team's job at a hundred thousand foot level is, you know, if you think about products or services like Elastic MapReduce as a product. We call it services here and we are the bridge between the service team that's developing the product and the field and my team packages up big scalable motions. Yeah, as well as first call decks. So we train and then, you know slash enable the field as well as specialist sellers on how to position EMR for example. And then we have what we call scalable motions or sales plays different firms have different terms for it. But like if somebody's running for example on-prem Hadoop, how do they migrate off of that and onto EMR? We actually have a program that helps spell out how to talk about the solution, a little

bit of the history, how to talk about the solution. What are the key assets both marketing, technical, etc. Total cost of ownership. So that way, you know, not unlike a Bain business, right? When you think about it, if you're my customer and you're saying hey *, I'm running self-match Hadoop. Can you help me? It's much better if I come and say yes, we've done this a lot. Here's our methodology. Yeah, you know that immediately instills, it's a consultative approach which immediately instills confidence versus yeah, we can absolutely help you. Let me figure out how to do this. And so that's what my team packages and does.

Interviewer
Wow, great to meet you Bob. Yeah. Okay, I will immediately share my screen then we can deep dive more into the material. So you should be able to see my screen right now. Yep, got it. Okay, great. I will give you a short background on these frameworks. So what we see over here is like the inter-decision framework and I identified three core decisions you need to consider when you want to implement your data product. So the data product from the Zhamak Dehghani book and all these decisions are based on grey literature sources. And what I mean with grey literature sources is everything that hasn't been officially published. So this can be YouTube videos, posts on Medium, white papers, blogs, things like that. So it's really from the practitioner point of view. And yeah, we can first briefly talk about the inter-decision links over here. Then I will go over each decision separately and then we can go back to this one and check if we're missing something or if something can be rephrased, things like that.

Expert k
And so let me ask you a quick question. So how this would work is you as in your capacity as an ISB / Bain employee, when you're talking to a customer that says, hey, I want to develop, I want to make better business decisions using my data analytics, machine learning, etc. This is the journey that you would walk them on.

Interviewer
Is that the easiest way to think about it? Okay, yeah, this is meant to provide some guidance to help you, yeah, to help you generate some kind of overview of what is possible, what kind of options you can choose from. So the first one could be what type of data product can be developed. So this is a bit conceptual, but it's really important for the other decisions you need to make in the future. And the second one over here is which approach is chosen for the creation of your data product. So do you want to start with greenfield development? So starting from scratch, or do you want to migrate from your legacy architecture? And this decision is about the outside layer. So it's like the infrastructure layer. How does our data product communicate with other data products, with self-serve platform, with the government layer, etc. Yeah. Then we have a decision on the product anatomy. So what's going inside the data product? What architectural design decisions do we need to consider over here? And what are the elements of a data product interface / contracts? So if we think about the perimeter of the data product, what kind of boards do we need on top of that? And the last decision is, of course, on deployment. So how do we want to deploy our data product? So if we consider the first one, this is about the data product type. And I actually identified three core options. So on the left, we see the decision. And here on the right, we see the three different options. So we can expose our data product as raw data. We can expose our data product as derived data. And we can expose it as an algorithm. And there are two deep dive variants on this one. So the algorithm can be an optimization-based decision support system, which is more like a BI tool. Or we can choose to have some kind of AI/machine learning model. And what I mean over here with the hybrid data product is more when we, is more a situation where we want to expose raw data as well as derived data, for example. So with that, we have two output boards

on the same data product. Yep, yep. And the composite data product is more when we want to merge something. So we want to merge the derived data with the algorithm. And we want to expose both of them as some kind of merge variant. But let's focus primarily on these three options. So on the raw data algorithm and derived data, do you resonate with them? Or have you-

Expert k
Let me go back to the first slide, the slide number two, and then come back to this in a second. Yeah, on this one. So when I talk to customers in a variety of capacities, because we have what we call, when we talk about our executive briefing center, what we call EBCs, right? They come from a myriad of topics. We have a topic called modern data architecture, and we have a topic for learn from Amazon. And so, and I used to talk a lot, obviously, about AI ML, right? And so the one thing that, and what I'm about to discuss impacts this slide, maybe, maybe not, and will impact the next slide. So just food with us. One of the first questions we have is what are the business objectives that they want to get out of what they're doing? Which maybe you handle that separately differently. But that may inform what type of data product is being developed. Yeah, I agree. What's their business objective? Number one. And then the other thing is what are their current hurdles to reaching that objective now? There are a number of companies that struggle with a data mesh and or data as a product. Yeah. And some of those challenges are technical, but just as many are organizational. Yeah, the social aspect is really important here. The social aspect is huge. So, and the other thing that's huge is making sure that you have executive sponsorship for this journey. And that's why I ask what's the objective that you want to get out of this journey? So, for example, do you want to do this? Do you want to unlock the value of data by driving additional revenue, maybe additional revenue per user, maybe a better personalization, maybe it's a new line of revenue. But if you don't have an executive sponsor, what will happen is you'll get down to the technologists, which are all good people and well-intentioned, but it could end up being a cost center, which is a problem, B, a science project, which we see a lot of machine learning projects turning that way. And so at some point, when you think about this journey, there's got to be, what is the business objective they're trying to meet? And the other thing is, so that's one thing. Then the second thing is, as you look at this, the other thing to consider is what are the current skill sets of the technologists that we're going to ask to implement this? And what do they want those skill sets to be in three and five years? And I segment that out because I've seen customers make architectural decisions because they say, you know what? I know my team needs to learn this skill, but we don't have time. So we're going to pay an ISV a ton of money, even though that we're going to change the architecture in two years. Very interesting decision, by the way. I've had others that have said, you know, *, we really want to adopt an advanced machine learning platform. And then I talked to them and they've got a bunch of folks that are Airflow experts or ML flow experts, and they're very much open source advocates. Yeah. So I'm like, well, your strategic decision may be undermined by the people that are open source advocates if they think they're going to be out of job in three years. And so think about your staffing model and hire for the future. And that'll impact some of these decisions. Yeah, sure. If I may say so, if I break it down to the first phase, second phase, third phase, fourth phase, in the second and third phase, you may realize that there may be some realistic constraints, like, hey, I want to do a transactional data lake, but I have nobody that knows Hootie and Iceberg. So are we going to staff up and will that take three months to get them on board? Or are we going to use an ISV solution or will we hire an SI to implement it? Like that, you may find some, I don't want to call them artworks, but you may find some artificial constraints, but you may find some constraints based on skill sets and or you may have to have an entire L&D learning and development program established.

Interviewer
Yeah, because these frameworks are really based on the architectural design decisions and not really on the social aspects. So that's something we lack over here.

Expert k
Yeah. Well, and that's why I just want to highlight it because I don't disagree with the flow, by the way, and we'll dig in more. But when I talk to companies, it's not a lack of technology choices and it's not a lack of desire. Generally, it's a lack of institutional alignment. It's a lack of knowing what their business objectives are, or it could be improper skill sets. So for example, I had one chief data officer, CIO for a FinTech firm. He's like, *, I am all about AWS and I hired some former Amazonians here and they're telling me that we need to use this ISV platform and not some native AWS services. And the reason is because we don't have enough folks that are comfortable inside of the AWS console. I think this is a mistake long term, but we're probably going to sign up for a very expensive license while we train our folks on AWS, which is a very interesting discussion. Yeah. I had another company talking to me about a similar journey as you're describing here, like, hey, we want to do a data mesh. We recognize that we have a lot of products that are in silos in the different businesses. The company they're part of has been the new company is basically a combination of mergers. The new CEO came from one of the mergers and he's like, look, we really want to do this data mesh, but I have divisions based in South America, for example, Latin America, I should say, that don't want to share their data. Yeah. So local strains, you mean. But it's not even, hey, are they set up on AWS or are there data governance restrictions based on the EU laws or Latin America laws? It was a, well, this is my data. Why should I share it with you? Yeah. Yeah. And so anyway, so I think as you think about this, I would not make light of the organizational dynamics.

Interviewer
No, sure thing. No, sure thing. Yeah, this is really focused on just one aspect of data mesh. If I had to cover everything, it's not the thesis. So it's just for a master thesis.

Expert k
No, no. Agreed. So I go to slide three. The other thing just to think about is, so I do agree that customers, and when I think about how they use the data and their business objectives, right. To your point, it's not uncommon to say, like, if I'm sharing data with you, and it depends on how you set up your data catalog. Let's start with that. And how you set up the data catalog, you could set it up to, like, I want to give you authorization for my raw data. Yeah. Or I want to curate the data, and I'm doing some experiments on it. And so now this derived data or the derivative, let's say, there's a subset of the data that I'm going to share with you based on a specific business intent or something like that. Yeah. So that I fundamentally agree with. I would suggest where a slight change possible to this is. Okay. A lot of times the derived data is best used for the AI model. So, for example, I may be doing something where I'm doing, you know, maybe I'm doing a customer 360 and you're in a sister division of the same company, and you're like, hey, I've heard * got this cool ML model. I want to get not only his data to see what he used and what's the features and how he configured that data, which is the derived data, but I also maybe want to share his model and see if his model is a good fit or take his model and retrain it on my data. And so what we see is a lot of folks specifically using derived data to expose into an ML model.

Interviewer
Yeah. Yeah. Okay. That's a very interesting point. So I can perhaps create a new arrow over here. That's what I would think. Yeah. Yeah. Good one. And do you think there is another

option to expose your data or do you think the tree?

Expert k
Well, I think, I think, look, I think that to your point, the exposing the raw data gets into again, it gets into now I'm going to expose my raw data to you. Once I expose my raw data to you, my assumption is, you know how to use that data. Yeah. Generally, you're not going to use that data raw. You are going to do something with that data. You're going to go in and look at the features. You are going to possibly augment that data with other data. You know, you may look at the features and say, hey, listen, delete some, get rid of others, do joins, transform it, you know, all to then, are you going to use a BI tool for it or are you going to use machine learning? I think the only thing that's sort of missing here is what is the on the middle line where it says exposed data product as an algorithm? Yeah. Got it. Then that has two outputs, which are logical. One is you're going to expose it as a decision support system, like a BI system. Okay. That makes sense because you've already, you've already run algorithms on it. You've already done something with it. Yeah. Whether or not how much that's used, right? The exposed data, exposed practice of raw data, that isn't an end goal of itself. If you take, if you take, yeah, yeah. So, if you take your, yeah, exactly. If you take your data as a, if you take this too literally your diagram, that's like, that's, it looks like it's the end result.

Interviewer
Yeah. Yeah. That's a good point as well.

Expert k
Just to make sure that a practitioner is not exposing data product as raw data to the consumer, but it can be exposed to consumer line data products. We will take it. And the other thing, all three of these will be predicated upon first to, what is the level of number one trust between my org and your org? Like you do, and then organizational skill set slash competency. Yeah. That's right. So, so like, for example, you using derived data or, well, you would expose your, your second one exposed data product as an algorithm is fine to be consumed by a business analyst that's using a BI tool on the back end. Makes 100% a lot. Yeah. Right. Or you, you, I don't know, like how many, I question, let me put it this way. How many data scientists would say, Hey, I'm going to take the data, expose an algorithm, and put it through a whole new ML model. Yeah. What, what's more likely as a data scientist would either take the raw data and, or the derived data and develop models. But it depends on, again, whoever's consuming that data must have certain skill sets to match these, these three buckets. If that makes sense.

Interviewer
Yeah. You mean to make your data product more, more multi-modal let's say so that we can.

Expert k
Well, to meet the end result, whatever the end. So data, the goal is not just to provide data. No. The goal is to innovate faster or to drive more value out of the data. And, and, and who does the other, the phrase we use at *, it's who are the personas that are consuming the data in each of these that will influence the next stage.

Interviewer
Yeah, sure thing. We can perhaps quickly go on to the next framework because.

Expert k
Yeah. Is this helping? I know I'm touching on the technology as well as the, as well as the soft

skills.

Interviewer
Good thing. It helps me to think more thoughtful about each decision. Well, especially about this one, because I think this is more conceptual, but really important. So yeah.

Expert k
And the only other one, I, the only other one I would pressure test is I would talk to people in your firm. People in your firm, if you're a data scientist, are you gonna take data just out of an algorithm and plug it right back into an ML model? Cause data science is like a play. Yeah. I could see that I would see instead of that exposed data being where it is, it's more on derived data or raw data. Yeah. There's a step in between the raw data. Don't get me wrong, but to run one algorithm on it and then, and then basically do it in an ML model on it afterwards. Seems like redundant.

Interviewer
Yeah. There could be an option in between. Yeah. Right. And the next one is more about, yeah, this is more about a business strategy. So do we already have a legacy architecture and do we want to migrate towards a more data mesh type of thing? Or do we want to start from scratch and really prefer Greenfield development? And the top four over here are more in terms of patterns we can use during migration. So for example, master data management, I think most of the data management is really important if you have a federated structure and there is a customer in your sales domain and a customer in your inventory domain, it needs to be the same central entity, right? So we need master data management to make sure of that. Strangle Fig is really about decomposing your monolith. So picking out a service and make a data product out of it. So, and doing this slowly over time. So we pick one service out of our monolith at the time and have a federated architecture. Yeah. At the same time. So it's really about decomposing and building up your data mesh architecture at the same time. It's more, yeah. Do you know what I mean with that?

Expert k
A little bit.

Interviewer
I can show you a picture quickly. That picture is always easier. Let's Google it. Strangle Fig pattern. Yeah. So this one explains it really well. So we pick a service out of our monolith and we make a data product out of it and we have our monolith at the same time.

Expert k
Oh, got it. Got it. So you basically are peeling off. Yeah. You're peeling off data. Exactly. Yeah. Yeah. Got it. That makes sense.

Interviewer
Great. Okay. Let me quickly go back to the slides. And zero trust. Yeah. We need to make the data products as autonomous as possible. So they need to have their own access controls. CQRS is about segregating your read and your write function. So those four patterns are, yeah. I think it's really important to keep in mind during your data product migration part. Do you agree with that or do you think we're still missing out on something over here? When we talk about migration?

Expert k
No, I think it makes sense. What I was thinking about is even when you do the master data management, like if you pick on that, you're going to establish your data. Like for us, you're going to establish... Somewhere there's got to be a master repository for your data and then all of the security, IAM controls, all of that on top of it, which is the definition of MDM, right? Yeah. And then typically, and this is where the blend between the first two. On the first one, you shift all the data and then you build your first use case to get credibility and then you build these use case. Second one, you move the data with each use case. That's the main distinction. We see that. I mean, we generally... I've seen more of getting all of my data into S3 first to use our technology just because they usually are trying to deprecate systems fast and do cost savings fast. So one thing as you look at these is the approach may be driven by business economics. Yeah. Right. And or strategic top down, you shall do X. All right. That sounds a lot like the Strangler Fig approach. So really taking it slowly and not starting right away with a federated structure, etc. Right. And now the approach is fine. But like in the current economic conditions, we have people coming to us saying, look, I'm on this legacy database or legacy Hadoop or legacy. And the first thing I want to do is get all my data out of this. Yeah. And later, like we see folks doing... They'll move, like, for example, Cloudera on prem to Cloudera on AWS. And then modernize their architecture by then going on to EMR. So specifically take a two phase approach in order to shut down their data center.

Interviewer
Interesting. And do you agree with, for example, the CQRS pattern over here? Do you think it's really important to implement this in your data mesh architecture? In your data mesh architecture as well?

Expert k
The CQRS? I haven't double clicked on that more for me. Explain that more.

Interviewer
Yeah, sure. Let me show you CQRS. So this picture represents it really well. It's about making sure that if something is read inside your data product, there are no... Yeah, nothing is overwritten. Immediately. So it's just to make sure that nothing is touched in your data product when something... When someone wants to read... It's like a read only access. So that's really about CQRS.

Expert k
Is that the end state or just a path? Because that... It's a pattern. So it's some kind of solution. So I generally see this as not mutually exclusive to the first two. So let me give you an example. I agree. Some of what you just described, every time we talk to somebody about establishing... Every time we talk data lake or data mesh, governance is a topic. Now, whether governance, they define it as access to data, or they define it as rules and regulations and permissions. But what you just described and the picture you showed me on CQRS is layered on top of both MDM and Strangler Fig.

Interviewer
Yeah, so these options over here, you can choose all of them, perhaps. And it's not mutually exclusive indeed. So you can do master data management as well as zero trust. So...

Expert k
But I do see a bigger distinction, just so you know, though. Okay. I do see master data man-

agement as being separate from the Strangler Fig, as being separate from Greenfield. Okay. Now, the question on Greenfield is... Like, I'd want to understand... I'd want to double click on Greenfield more, just because if you think about it... And I'll just tell you about patterns that we see. So I'm trying to give you examples that might fit and support your framework and or make you just challenge it and question it, right? So we have customers that come to us and they... Like I said, hey, I'm on... From an AWS standpoint, the first question is, where currently is their data? Is it on-prem or is it... Is it outside of AWS or is it AWS? If it's AWS, and maybe it's a combination of AWS accounts spread out, right? So now the question is, okay, we've got... Maybe they've got three different data lakes because it's a... Different lines of businesses in an insurance company, for example. They've got their auto insurance data lake and they've got their homeowners data lake. Yeah. And so the question is, how do you build out that federated data mesh? You may not elect to do master data management. You may not elect to say, hey, let's hydrate a brand new S3. Yeah. You may not migrate them anywhere, not this Strangler Fig. That may be what you're calling the zero trust architecture where, hey, we're going to keep our server data lakes, but we're going to do a data mesh. And we're going to have an agreement set up both architecturally as well as philosophically between the consumers and the providers of data. Yeah. Right. And so the interesting thing is on both of those, whether you're coming off of outside of AWS, that could equate to both the master data management as well as the green field, because, hey, the first thing I want to do is shut down my data center. So I'm going to pull all my data and drop it into S3. Once it's into S3, I'm going to layer all the permissions on top of it, make sure we've got the governance set up, and then set up my next, my first set of use cases. And then I'll start doing, you know, I'll query the data, maybe I'll do some feature engineering, and I'll set up derivatives of it and then run my machine learning.

Interviewer
Yeah. Yeah, I agree. Really interesting though, that you can do greenfield development and of course, most of the data management at the same time.

Expert k
It all depends on where they're starting from. Yeah. Right. You know, because, and you know, because if there are, like I said, if they're already all on AWS, I'm giving you the AWS view because I'm like, that's what I see. Yeah, sure. I think the challenge that you're going to run into is, let's say, you're dealing with a EU firm, and they're very happy running their own data center. Yeah. Now the question is, okay, master data management, but you've got three separate data centers, one in France, one in Germany, one in the UK. All right. Do you do a federated data mesh across those zones, or, and maybe you do, and maybe you even do replication of data, or fault tolerance, which is frigging expensive for them, but it makes sense, I guess. Or do you say, look, no, we're going to do master data management by, we're going to have one data center, we'll hydrate everything in there. You know, but so I think to a certain degree, where they're coming from, their point of origin will influence some of this, but not all.

Interviewer
Yeah, sure. That's also about the third framework, actually, which is more about the infrastructure. So what we already have and what we have around our data product and how resources help the data product as well. Really more extensive, to be honest. Yeah, the first option to consider is, of course, a schema registry. So when something changes in our internal storage, in our data products, internal storage is changed, is being published in a Kafka topic, for example. And all the data products that are subscribed to this Kafka topic will get notified. But the change event isn't that user-friendly to read. And that's where the schema registry comes in.

This schema registry makes sure that this change event is converted in a more user-friendly format. And I think we already talked briefly about the central data product catalog. So this is more like a global catalog keeping track of all the metadata in your data mesh. Now, event streaming backbone. And this can be Kafka or Kinesis, which is the Amazon one. The shared storage. Some practitioners opt for shared storage because if you have like 10,000 data products and each data product would have its own internal storage, you need to run a federated query to search something in your data mesh. And this can be very costly. But on the other hand, if we choose for shared storage, it makes our data product less autonomous. So this is some kind of trade-off we need to make. So do we want to look more from a master data management perspective, or do we want to look more from a federated perspective? So the data mesh perspective. And API invocation is about, yeah, how can we expose our data product to the consumer? So for example, with the GraphQL, gRPC, REST APIs, but we can use a SQL access point as well. And on the bottom, we see non-functional requirements. So of course we need security controls. We can implement an in-memory cache and we can think about a query catalog. And what I mean with a query catalog is that you have some kind of manual for your data analyst to help him or her with their queries. So it's like a manual where all the possible queries are defined for your data product. So this is really about the layer on top of your data product to help it communicate with the other parts of your system, of your data mesh. Do you agree with most of these?

Expert k
I do. Again, I get to, there's a, they're not 100% mutually exclusive, right?

Interviewer
No. So. No, it doesn't have to be mutually exclusive. It's just to provide you with an overview for, like, I can choose for this one. I can choose for central data product catalog. So you can choose multiple options over here.

Expert k
So here's, so let's say, let's say, let's say, let's say, here's, so like, let's say you were somebody that worked for me. Let's say you're on my team. What I would tell you is for slide four and this slide, what would ground it in probably inside of Bain, as well as with your clients. Give me an example, give me a real customer example, even if it's anonymized. So for example, I see the slide and, and in an ivory tower, and please don't get offended by that, but theoretically, I get it. Now ground it for me, give me an example. So like, like you may want to discuss, hey, we've got, a media entertainment company that is streaming data to their customers. They've got a shared data lake for all of their products, right? Yeah. They're using Iceberg as their table format. So it can be a transactional data lake and they can keep, keep track of things. And they're using, you know, MSK and or KDS probably with some KDF to it, to look at, okay, what's the real time events that are happening, you know, publishing them, making real time decisions, recommendations of its streaming media, let's say for what they're watching. And they're bifurcating that, that streaming backbone, it's event streaming. So it's going to be number one, you know, branch it immediately to run your ML models for predictive for, for suggesting the next movie and then hydrating it back into your central data product catalog or your S3 for future models, for future recommendations. Right. And so what, what would help when you show this slide is the very next slide to say, now let me walk you through a customer and how they thought through what I just showed you and what their architecture looked like. That would, that would allow your audience to get grounded. Does that make sense?

Interviewer

Yeah, sure thing. So to use use cases as an example of how, for example, the schema registry works and.

Expert k
Yeah. Use a real life customer, use a real life customer example. Yeah. That either you have seen or some or one of your peers has seen to say, now let me. So listen, I talked to you about this framework and I talked to you about the different products and how they interact with each other. So let me go from, from the, from this great concept. And let me, let me describe to you a customer that is doing this real time. Like, yeah, if you watch a football game powered by a, they were stats, right? They've got a data lake, right? That has all the statistics about previous plays, et cetera, et cetera. So as they're running a real play and, and you're streaming that data, you're saying, oh, this play matches this data. Okay. It's got a 35% probability they're going to score. Yeah. Right. And so how, and so that uses the central data of S3, you've got the schema registry of who's changed what data over time you're streaming real live, what's happening in the game. You're running an ML model. And by the way, you know, you've, you've got privileges, privacy, governance all over cascading over it. That's just, sorry, work never stops. That's just one way to sort of ground it. Yeah. That's a good one. Yeah. I should definitely have a look at that one.

Interviewer
Yeah. We can perhaps quickly go on to next or, because I want to be cautious of your time and we only have 12 minutes left.

Expert k
Yeah. Let's go on to the, well, it's up to you. I I've got about five more minutes for, I do have to take a short break for my next meeting, but, oh yeah. You can either ask me any other questions you have, or we can go on to the next one. Whatever would best meet your needs.

Interviewer
I really want to discuss this one. And this is the last, actually the last one I want to discuss this one, I think is really related to, to the data product and that, so it's not really important to discuss. Okay. So yeah, let's quickly go over the data products anatomy. So what's going on inside the data products? Well, if we choose for event streaming, a change data capture is really important because this change data capture captures changes in your internal storage. So if something changes, this one will know, and it will convert this change into a change event and publish this in the event streaming backbone. The immutable change audit log is, can be an S3 buckets where we store all the changes that happen inside the inside the mesh. So if something is changing, the change data capture will notice it will publish this in the Kafka topic, for example, and all the immutable change audit logs in all the other data products. These will store this change. It's like an append only mechanism. That's what, that's why it is immutable. So it doesn't change over time. It's appended in the, in the store.

Expert k
It's almost like snapshots of the history.

Interviewer
Exactly. Exactly. So that's really important. So everyone in the mesh is notified about your change in your data product. Internal storages, I think we discussed this quite a bit already, and that can be again, a data catalog. So this is more like a local data catalog that is really specified on your data product and not on the global mesh, let's say. Yeah. And it can be an

observation plane on your data product to make sure that the data scientist is able to observe the data quality of your data products. The control plane is for your government team. And this is really to enforce certain global policies in your data product. The data onboarding, yes, this can be EMR, for example. So all these spark processes, all the transformations that are being applied to the data set in your data products. So yeah, this is a really high level perspective of the architectural elements involved in the data product anatomy.

Expert k
Yeah, it makes sense.

Interviewer
I think I covered quite a bit of this, but do you think I'm still missing out on something considering architectural design decisions?

Expert k
No, I think it's okay. I think you've got most of it.

Interviewer
Oh, great. Because the next one is indeed about the ports. So the observation port is connected to the observation plane, for example, the control plane. The control port is connected to the control plane to enforce your policies. The discovery port is connected to your local data catalog. Yeah, to get some kind of sneak preview of the metadata inside this data product. And some practitioners mentioned that I forgot the two most obvious ones, so the input port and the output port, which is really important for your data flow, of course. So yeah, I think there can be five data, well, five ports on top of your data product in total. Do you agree with that, that we have just five ports and nothing more? Or can there be different ports as well?

Expert k
No, those are most of them. You have the control, observation, input and output, publish and as well as consume. Yeah. The only one to think about is, and maybe it's separate from consumption, which is making sure you're taking snapshots of the data. It's less of that's the only, you probably got that thought and it was on the previous page, but that's the other one.

Interviewer
Oh, that's a good one indeed. Yeah. And actually the last framework, I definitely have to reconstruct this one because it's a uses, first of all, provide their names. I shouldn't use that. This one is about containerization using a container orchestrator, but we can of course also deploy our data product by using serverless functions. So this, you should interpret this actually as a containerization and serverless functions, but I'm sure there are a few more options to consider when you want to deploy your data product. Do you have any idea on other things we can use during deployment?

Expert k
Well, so there's, there's, I don't want to say bare metal, but like, if you think about S3, right, you know, folks, so we see, we see some companies that do certain workloads on Kubernetes and others not. So EMR, for example, which is our big data processing engine. I don't like calling it a compute engine. Some people do. It can run on, it's got, right now it's got three deployment patterns and they are totally separate. They can run alongside each other, but it's either Kubernetes or it's serverless, which is sort of on EC2, but there's no long-standing instance. Okay. Or it's on EC2. So when you start thinking about deploying a data product,

for us, those are for my biggest analytics product, those are the three ones, Kubernetes, EC2 or serverless. Now serverless is still using EC2, but it spins up when it's needed and it shuts down when it's not needed. Right. So you mean that I should include some kind of computing layer? Yeah, because if you think even if you're using a serverless, you're not going to use Kubernetes. Even under Kubernetes, there's compute underneath Kubernetes. Yeah. Right. And so what if they elect not to use Docker and what if they elect not to use Kubernetes? Yeah. And I mean, this slide works, but it's Docker slash, this is container centric is what you've done here. Yeah. This is only one option indeed.

Interviewer
Yes. When we're talking about compute, I immediately think about storage as well. We need a storage layer.

Expert k
Correct. But with storage, I don't see folks running, like S3 doesn't run in Kubernetes. It's a different thing. So if this is your compute layer, let's start with that. Yeah. And I'll take, let's take Elastic MapReduce, which is whether you call it Hadoop or whether you call it EMR, which is our product, we generally see three patterns for deployment and they coexist. So I have some very big financial services firms that are running long running clusters of EMR, because they've got really big jobs to run. Yeah. But they're using EMR serverless for data pipes. Okay. Right. I've got other ones that are saying, no, I'm not going to do serverless. I'm, you know, we, and I'm not in agreement of this, by the way, but they'll say, hey, we are a multi cloud environment. We view Kubernetes as our extraction layer, so we can go across different clouds. Yeah. Therefore, I'm going to run EMR on Kubernetes. I've got it. Yeah. So that's the big on this one. If you were to show this to somebody, that's anybody that runs on the cloud, let me put it that way. Let me phrase it differently. Whether it's us, Azure, or that's the first thing they're going to say, which is, yeah, look, I'm either on, you know, I don't need to be on Kubernetes. I could really run on Azure core itself, or I could run on EC2, you know, you know, I can run on my basic compute layer.

Interviewer
Yeah. All right. Yeah. I definitely think this needs some reconstructing. Yeah. Let's have only two questions left. So if we go back to the first one, this is, these are all the decisions we just discussed. Yeah. We already briefly talked about the business strategy. That needs to be your first decision actually. So besides that, do you think there can be more decisions we need to cover in this inter-decision link framework? Or besides the business strategy, it pretty much covers everything in your data product implementation.

Expert k
I think this is pretty clean. I like this because you're not making it complex. Some of your other slides get really complex fast. Yeah. Indeed. Yeah. You know, this is fairly clean. What's the, I mean, again, I think the phases are right. Yeah, I think this is fine.

Interviewer
Yeah. And do you think it can help users of this framework to, to, yeah, to implement a data product? Do you think it can be useful or?

Expert k
I think it, I think what you'll need to do is I think for all of these, you're going to get similar questions. If you were consulting me and I was a CIO, I would say, yeah, a lot of this is

intuitive. It's not all mutual exclusive. There's definitely relationships and you'll have to say, yeah, I think an example, like I would take a real customer example. And I would say, Hey, let me talk. Let me tell you. Like, okay. Sorry. I've got to, I've got to, are you still there? Right. I've got to hop on this call. I think we're out of time.

Interviewer
Oh, no worries. Yeah. I would ground this with a customer example, but let me just say, I think it's a good example. I think it's a good example. I think it's a good example. I think it's a good example. I would ground this with a customer example, but let me drop and get this call. Yeah, sure. Thank you so much for this interview, *. Yeah. Hope to see you soon too. Bye-bye.

Expert k
Bye-bye.