1:27

Alexey

**This week, we'll talk about data access management. We have a special guest today, Bart. Bart is the CEO and co-founder of Raito. Bart is on a mission to give data workers access to the data they need, to do their job in a faster and safer way. Before co-founding Raito, Bart worked as a senior product manager in the data privacy area at Collibra, and Bart believes that for data to be liquid, trust has to be solid. He will probably tell us later what it means for data to be liquid. [Bart agrees] Because I have no idea. [chuckles] My imagination is not good enough to understand this metaphor. So welcome to our show.**

2:09

Bart

Hey, Alexey, thanks for having me. Super excited to be on the show today.

2:14

Alexey

**As always, the questions you will hear today were prepared by Johanna Bayer. Thanks a lot, Johanna, for your help.**

# Bart's background

2:23

Alexey

**Before we go into our main topic of data access management, let's start with your background. Can you tell us about your career journeys so far?**

2:33

Bart

Yeah, Alexey. That was a really great summary before starting. I was, indeed, a product manager of privacy at Collibra. That's where we also saw the challenge of scalable data access methods. Now, Collibra was my first step in the world of data and software. Before that, I was actually a consultant at Deloitte, where I was doing financial risk management. But there, I saw that the future is data – that everything revolves around data. That's why I made the move to Collibra. But I learned that for data to be really successful, and for organizations to be really successful, they need to have that trust. And that's why I started Raito together with Deiter, my co-founder.

3:26

Alexey

**So what did you do as a consultant at Deloitte?**

3:31

Bart

We helped banks with accounting, with valuation of financial products – there was a lot of mathematics, a lot of physics, some models that we borrowed from physics. But then, gradually, I was also getting more and more involved in data governance programs because this was right after the financial crisis. That was where, for banks, it was impossible to know how much loans were defaulted, and what their exposure was to other banks. Then the European regulator actually boasts data governance regulation on banks to improve data quality, literature, and so forth. In that way, I kind of evolved into the data governance space.

4:20

Alexey

**Basically, there was a financial crisis and banks had some data, but it was a huge mess. And what you did as a part of your job as a consultant at Deloitte was help them make it less messy. Right?**

4:40

Bart

Indeed. I mean, talk about trauma, right? This is before you had all the hip data governance solutions that are out there now. If you have to do data governance now, it's much more pleasant than when I had to do it. I had to draw and manage it in something like the MS Office tool. I had to keep data dictionaries in Excel. So this was before all the tooling was there. And I can tell you, it was not great. By the time my lineage was ready, it was already outdated. By the time my dictionary was ready, it was already outdated. And I had to go back.

# What is data governance?

5:20

Alexey

**So what is data governance, actually?**

5:24

Bart

Data governance is... each time the term “data governance” is mentioned, you see people frantically looking for the nearest exit to escape the conversation. It's a heavy-loaded term. If I would have to summarize data governance, it's just all the activities that you do to create trust in the data that you're using. It's that your data workers – your data analysts and scientists – can trust that they're using the right data. It's that they can trust that the data has sufficient data quality. But it's also just that your customers can trust that their data is being used in a safe way.

If you put it in that perspective, who doesn't want data governance, right? So despite the bad connotation that data governance has, if you're really serious about using data as a competitive asset, you need to create trust, and hence, you need to do some of those activities. But the cool thing is that the perception is shifting. You have all these cool startups now (among others) that help you do data access or data governance in a more scalable way.

6:43

Alexey

**Why does it actually have a bad connotation? Why do people search for the fire exit when you talk about it? [chuckles]**

6:52

Bart

It's because of how we used to do data governance. It used to be very centralized, right? A bit like in an ivory tower...

7:02

Alexey

**Like those Excel sheets that you talk about?[chuckles]**

7:04

Bart

Excel sheets, yeah. So it's a couple of things. It's awareness. It was very hard, and it's still today, very difficult to make people aware in the organization about the importance of good data governance. But there's also the way that we used to do data governance. It used to be in an ivory tower. Your boss issued standards that he tried to impose on the organization, and it was always bolted on data products, bolted on how we use data. That creates a lot of friction, right? It makes it that a data governance team is more policing.

7:40

Alexey

**But why does that happen? Is it because somebody from... For example, in your case, when you worked at Deloitte, the EU commission or some governmental entity said, “Okay, you have a mess, banks. Go fix it.” And then top management in a bank probably thought (or just told others), “Hey, we have a mess in the data. Go fix it.” So this was a top-down decision and it was forced on people. Plus it was too difficult to implement. Right? Is that why it has this bad connotation?**

8:16

Bart

Yeah, indeed. We're learning. Data governance is going through a transition. How it used to be was, indeed, top-down and forced on the organization. People had to do it and they didn't want to do it. But what's happening now is, it's becoming more and more grassroots (bottom-up) where data governance is being implemented – in the DataOps process. Data engineers can integrate in the development process, making it so much easier and it becomes more convenient. Also, they really see the value from it. So in that perspective, it's really evolving in a good way.

# Data dictionaries and data lineage

8:58

Alexey

**So you mentioned a few things. You mentioned a thing called data lineage, where you use Office tools to draw. Then you mentioned that the data dictionary was in Excel. But what actually are these things, and why did banks care about them? Why was it important for them to actually have these things? Why did they care about that? How did it help solve their problems?**

9:24

Bart

I don't think they really cared back then. This was 2009-10. It was basically the BCBS 239 regulation that kind of imposed a better process (better data governance) on banks. They had to comply. That's why it was top-down, because you had to do it. Nowadays, people are more and more becoming aware that you have to do it if you really want to use data. That's a bit of the mind shift that's happening today.

10:02

Alexey

**And this lineage and data dictionary – what are they?**

10:07

Bart

The dictionary, or the catalog... let me start by establishing what a catalog is. A catalog is just a list of all your datasets that you have in your organization. You just read the description to see what you can find in a dataset, and also just the definition. For example, “customer” has a different interpretation depending on where you are in the business, right? In finance, a customer is somebody that's paying the bills (the invoices) whereas in sales, a customer is somebody that is becoming a customer. Different definitions. You have to be sure what the definition is if you look at the report. And lineage is a traceability view, or a view of how your data flows through your application and data sources.

11:00

Alexey

**So it's like “Okay, this data starts here, then there is this transformation applied to this data, and this transformation. And this is how it ends in a report for the top management.” So you can kind of see the entire journey of the data.**

11:17

Bart

Yeah. That's correct.

# Data access management

11:20

Alexey

**Okay. And what does it have to do with data access management? What is data access management?**

11:27

Bart

Yeah. Well, data access management is a weird conundrum. It's security – data access management is really like a form of access management. It's security. But how it has historically grown is that it's now a responsibility of the data governance teams – of the data teams. In this way, it's a bit weird. [chuckles] But to understand data access management, you've got to take a step back. Access management used to be... we used to manage access at the application or database level, You had access or you did not. But once you had access, you had access to all the data in there.

But now, cloud computing and cloud storage have completely revolutionized the way that we use data. What you see is that data teams are now massively building analytical workflows or data science workflows on a public cloud. By doing that, they're taking data from all these different systems and applications and moving it to one location: your lake, your warehouse. And by doing that, they lose the Chinese walls that they naturally had between the different source systems.

Now, we have all data in one location, and it is being publicly accessed over the web, of course, you need to have the same Chinese walls in your lakes and the warehouses. You've seen all the data breaches – you look at LinkedIn or a newspaper, and there are data breaches in there. There's all the privacy regulations. Your customer awareness is increasing. So you need to have these same Chinese walls in your warehouse and in your lake. That means managing access at the dataset level and that's data access management.

13:23

Alexey

**Okay. You mentioned that data access management is the responsibility of the data governance team. So what actually is a data governance team? What do they do?**

13:34

Bart

Well, it kind of depends, right? What we see is that in organizations that are really data-driven of up to 3000 people, their data access management is still the responsibility of the data team – the data engineers. Then for larger organizations, you'll see that they start building out a data governance team who will be responsible for the data access management. Or in certain cases, when they're implementing data mesh, then it's the data owners who are responsible for data access management.

As for a data governance team, the nature of that team is kind of shifting. It used to be dictionaries, definitions, the lineage, the catalog, and so forth, but now it's also becoming more and more of a team that promotes data usage (does evangelization) and enables teams to use data in a more efficient way. So it's going more from monitoring and policing to enabling the business.

# How to learn about data governance

14:47

Alexey

**So in smaller organizations of up to 3000 people... very small organizations, right? [chuckles] In these organizations, there is a team (I guess a central team) of data engineers that moves data from one place to another. As a part of that, they implement some of the things we discussed. Right? So it's just data engineers who do that. But how much...? I'm thinking of data engineers whose core skills are cloud, building all these ETLs, Spark, Python, Docker – all this stuff. Very technical people.**

**How do they actually know about what they need to do? Do they need to educate themselves in data governance and data access management? Because it doesn't seem like their core job. Maybe it should be, but it's not – at least from what I typically see data engineers doing.**

15:49

Bart

What you've just described is a core problem. Look, if you're a data engineer who is interested in data governance, please investigate. Please learn more. It's only going to improve your career. But it's not your core competence, right? You haven't studied, or you haven't delved into data engineering to do data governance. That's a bit of that challenge that we're seeing, specifically for data access management as we're talking to a lot of organizations. The data engineers have to do it, because that's how it was – they get access requests, so have to manage access. And they have to do it because they know the technology. That's the only reason.

They're not doing it because they know the business context or they know the policies. No. It's because they know the technology and they're using it. So they end up being the people who have to process the access requests and that's not how it should be. Right? It's not their core competency. It's not what they were hired for. They should be focusing on other things, like managing infrastructure. And that's why you see that somewhat larger organizations, indeed have realized that, “Hey, this shouldn't be with the data team. Let's give it to the data owners or the data governance team.”

16:59

Alexey

**So in larger, or maybe more mature organizations, there is the central data team with data engineers. But in addition to that, there is this data governance team. Correct?**

17:09

Bart

Yeah. Or a data mesh team, where the responsibility is with the data owners.

# What skills are needed to do data governance effectively

17:18

Alexey

**And what kind of skills should these people in the data governance team have in order to be able to do these things effectively? Where do you actually learn these things?**

17:30

Bart

Concerning the skills, I'll ignore the data owner for the moment. In the data governance team, I actually think the biggest skill is... you don't need to know all the technology. You don't have to be super technical. You need some kind of business affinity. But you also need to be able to enable and manage change. To be successful, people have to change, processes have to change, etc.

You need to be a change enabler. That's probably the biggest skill they can have working in data governance – “How can I change behavior? How can I make people aware? How can I evangelize? In terms of hard skills (pure knowledge), there's the DMBOK (DAMA-DMBOK: Data Management Body of Knowledge). It's really like the Bible for data governance people.

18:33

Alexey

**Okay. You said the skills that they need is change management, which does not sound like something that engineering or computer science courses cover at any university I know. I guess these people do not necessarily come from a computer science background. So what kind of background is actually best? Where do they usually come from? Like you, from the consultancy space, or where?**

19:07

Bart

I've seen data governance people with all kinds of backgrounds, actually – engineering, legal, HR, marketing...

19:19

Alexey

**So it's not something you can learn in school. For instance, when you finish school, you can't say, “Oh! Let's go learn about data governance!” and have a Bachelor's degree in that.**

19:31

Bart

Not that I'm aware of, but I wouldn't be surprised if that comes about soon. [cross-talk]

19:38

Alexey

**I guess it would be a specialization in some sort of other field. Right?**

19:48

Bart

No, I mean, like I said, it's really fundamental. I was at Coalesce in London last year (a DBT conference). This is about analytics engineering. There were the keynotes. And after the keynotes, I was with the co-founders and the product manager of DBT. There was a Q&A, and half of the Q&A were data governance questions – they were data governance-related. These were analytics engineers, they were building on DBT. And because they were focused on proving the value of data – getting these initial data products out, those initial successes. Then they had success – they got more and more data consumers on the platform, and more and more data products in there. Then you start seeing all those data governance issues. And data access management is really one of those challenges.

That's why you see that the data teams that have proven the value of data, they're getting more data consumers on the platform, they're successful in their self-service analytics, the number of data products is exploding. And then they're getting all these data access management issues. It's the same with data governance – it applies to data access management. Now you've got to gradually mature your data access management, right? You cannot just say, “Oh, stop the presses. Hold on, everybody. Stop doing self-service analytics. I've gotta improve my data access management.” That doesn't work. That's too disruptive. That's also the approach for data access management. It should be done incrementally. Don't have a big bang, but incrementally improve your data access maturity. It's the same with data governance. That's where you need a good plan, a good change agent, that allows you to do that.

21:50

Alexey

**I understand. What about those poor data engineers working in smaller organizations who, all of a sudden, need to deal with all these data access requests? I assume some of them might naturally become interested in data governance. So how do they actually get better at that? Do they pick up the book you mentioned (DMBOK)? Or how do they do this?**

22:14

Bart

There's a lot of content out there. There are also several Slack communities. There's a data mesh Slack community. The name escapes me, but there's also a data governance Slack community where you can go and check out. So there are definitely communities and literature out there that can help you.

22:37

Alexey

**I think there is a book from O'Reilly that is called Data Governance, right? [Bart hesitatingly agrees] I think we had the authors on this podcast a couple of years ago. So if anyone is listening and did not check that out, maybe you can go and watch that. Just type “data governance” into our YouTube channel. We discussed it in much more detail.**

# When an organization needs to start thinking about data governance

23:03

Alexey

**So when is an organization large enough to actually start thinking about the data governance team? Or is it more a question of maturity? How does it happen?**

23:14

Bart

I think it depends on several factors. What I see in practice is that this is for organizations that are very serious about data. They have customer data, and they want to use it as competitive assets – for better insights, better services, better products, and so forth. Then you have to really have the basis of data governance to become successful at that. Now, this is typically when the organization is of a certain size. If you're small and can just talk to each other over the computer screen or over Slack, as we're working from home now – but as of a certain size, if you're serious about data, then you really need to invest somewhat in data governance. Because, as I said, it is the fundamentals to become successful with data.

24:16

Alexey

**Let's say, a startup with 10 people – do they need a data governance thing?**

24:23

Bart

No, I don't think at that stage you need data governance. You all have the same definition of the data, you understand the data flows. For data access management, just one person manages it. It's when you start getting more data consumers on the platform, when you start getting more data products in there, and you start feeling like the issues are piling up, such as people not trusting the reports anymore, data quality issues. You see the trust disappearing – the trust fading. You get a lot of ad hoc requests, put out a lot of fires. These are all symptoms of a lack of data governance.

25:05

Alexey

**Let's say if we have data issues, and we already lost the trust in data, isn't it too late now to do this? Should we have tried to prevent it?**

25:17

Bart

Yeah, indeed. The priority is often, “Let's prove the value of data first,” In theory, yeah, it should have been prevented. In practice, if you have a limited budget and your responsibility is to prove that there's value in data analytics, you're gonna focus on that. And when that takes off, you're going to improve your data governance. That's how it happens in practice, I would love to see it the other way around. But that's how it is.

25:44

Alexey

**[chuckles] So I guess for a startup with 10 people, it's like two teams working together at most. But let's say that there is one team that is producing data, and then there are multiple teams that are consuming this data, the moment you have some sensitive data that only one team can consume and others cannot, such as some personal identifiable data, then you need to start thinking, “Okay, how do we make sure that this team can have access to this data and this team cannot?” Then you need to start thinking, “How do we keep this organized? How do we not go mad thinking about all these things?” Because there could be another field and other sensitive kinds of information that maybe just one team wanting can consume, and it's a different team. Right?**

26:33

Bart

Absolutely. I think you need access management before you create a data governance team, that's for sure. As soon as you have a data warehouse or a lake, and you have sensitive data in there – be it customer data, or business-critical data, like very sensitive intellectual property – of course, there you need data access management. That's for sure. And then, like you said, if you start bringing on more data consumers, then you really need it. But then you need to do it in a scalable way. You need to have automated processes, better insights, better collaboration.

That's when you have to start looking at taking that responsibility and the processes out of the core data team, and then put it either with the governance team that typically starts developing around that size – you start getting your first data steward (data governance person) or push it to the data owners. But if anything, when you work with customer data, data access management is one of the parts of data governance that you have to start with as early as possible.

# Good data access management processes

27:49

Alexey

**You mentioned the word “process” many, many times. I'm wondering, what should a good enough process look like if we talk about data access management? Let's say we have producers of data, we have consumers of data, some of the data is sensitive, so not everyone should have access to that, and one team wants to have access to this data. How do they go about starting to consume this data?**

28:20

Bart

I didn't realize I was talking about processes that much. I must be getting old. [chuckles]

28:25

Alexey

**Three or four times. Yeah. [chuckles]**

28:28

Bart

Yeah. Well, there are many processes in data access management. You have the access request, approval, you have the regular reviews, you have revoking access, and so forth. I think it starts with good roles and responsibilities. Ideally, and this is for somewhat larger organizations, you need a good collaboration between the data owners who know the business context of the data – who know what it means. Ideally, they approve access, they manage access, and they review it. The data owners work with the data engineers – they know the technology. They have to implement the controls.

And then, if you have a data governance team, they should also be involved because they know the regulatory context within your work. They know the privacy and security standards, and the policies in the organization. Ideally, you have a collaboration between these three. What we've seen is that when data access management does not scale, it's because the collaboration between these three elements has broken.

29:36

Alexey

**I'm thinking that it's still a bit abstract to me. Let's take an example. For example, we have a use case like churn prediction. We're in a company, we have some clients – it could be something like your usual internet SaaS company. There is a customer team that manages all the information about the customers like their contact details and so on. And let's say there is also a marketing team that wants to understand if somebody is about to churn. If so, we want to try to win them back by offering some discounts. So the marketing team wants to implement this churn prediction. The marketing team needs access to the email field. How do they go about this if we have proper data governance and data access management tools? You mentioned that there are things like access request/approval/revocation. People probably ask the requests, right? How does it happen?**

30:40

Bart

Ideally, there's this concept of shifting governance left. Ideally, when the data engineers have created the initial dataset with the customer emails, they have decked that dataset with customer email, and they've also defined the roles that can access that dataset.

31:07

Alexey

**So that's the customer team, right? The data engineers in the team that make this data available, it's their responsibility to document this data and say which of the fields (which information) is sensitive, which is not, and who can access what. Right?**

31:23

Bart

Yeah, ideally. And that's when you start moving towards data mesh and data contracts. The data producers in a sales team say, “This dataset contains customer data, These are the roles that can assume it.” And then the marketing team basically, when they need that dataset, they find it in a catalog, and they say, “Hey, I need access to this dataset – the email field is needed for churn analysis.” So they log an access request with the purpose of 'churn analysis'.

31:56

Alexey

**An access request is when they say, “Okay, this is the field we're interested in, but we don't seem to have access to this. This is the reason why we need it – because we want to do XYZ.” Right?**

32:08

Bart

Yeah. Ideally, they also say, “We only need it for three months,” or “We need it permanently.” But the point is that you also set the time dimension. Then if you have a data owner in the sales domain, they approve it. Then, in an ideal case, permissions are automatically updated. If that approval is for three, four months – after that period, the permissions are revoked. That revoking access is really important. What we've seen (this is pure data access management, not data governance) is that one of the issues that we see with data access management is excessive privileges – role explosion.

It becomes way too difficult to see who has access to what or you have a lot of excessive users with excessive privileges. That's because access is always granted and not revoked. I think now, more than ever, with public cloud and all the breaches, privacy regulations, but also the upcoming security regulations in Europe and the same in the States, you at least need this concept of privilege access management. Time-bound access is also part of the best practices I would say.

33:21

Alexey

**You're based in Europe, right? In Belgium?**

33:22

Bart

Yes.

33:23

Alexey

**It's especially important to know about all these things and in Belgium, the European Commission is actually there. It's part of the Euro Parliament. So all these laws about data privacy are actually made somewhere near you, right?**

33:40

Bart

Yeah, those are my neighbors. [laughs]

33:45

Alexey

**Do you go to have barbecues with them?**

33:52

Bart

[chuckles] They're my neighbors. But my friends are more like other entrepreneurs.

34:01

Alexey

**Okay, I see. [chuckles] So I guess in this case, if I request temporary access – if I say, “Hey, I want to do a proof of concept. I want one to see if our churn prediction model is actually useful,” then you request access for three months. Then, if it's granted, whatever data access management tool we use should automatically revoke it after three months, right?**

34:24

Bart

But you have to look at the development cycle. Typically, data engineers or data scientists (data scientists in this case) have to build a new model. They request access to the data to build that model. They get it for a couple of months. When the model is built and can go into production, the data scientists don't need that data anymore. We give access to the service accounts, or the workbench that needs the data for the churn prediction model, while access for the data scientists becomes revoked. I'm always talking about privacy and security.

You have many regulations pertaining to why you have to limit access. You also have [inaudible] and so forth. But it also just prevents you from making mistakes. For a lot of people, having the guardrails that prevent them from deleting a production table or whatever, can also give peace of mind. So there are a lot of benefits in just working with temporary access and knowing that “When I do something wrong because I didn't have my coffee in the morning. I have my peace of mind knowing that is somewhat limited to the access that I have.”

# Data masking and the importance of automating data access

35:35

Alexey

**As a data scientist in the past, I can think of ways where it could potentially go wrong. Maybe there is a solution. For example, say we implemented a proof of concept. We decided to go ahead with this. It's in production right now and I no longer have access to this data because why do I need access to emails of the customers? But something happens with our model. It gets broken and then I need to go and debug it and figure out what's wrong. If I don't have access to this data, how do I actually debug it?**

36:11

Bart

Yeah. I mean, I don't want to go too much into Raito's solution, because we cover all that stuff at Raito, and I don't want to make it a sales pitch. But in this case, ideally, you get access to do your investigation. Currently, what is broken is the process of requesting, getting access, and revoking access. The process does not work. That's why, in your experience as a data scientist, I'm sure you've had admin rights. They were just like, “Here, Alexey. Leave me alone. Here are admin rights. do your thing.” Right? That's because the process is broken.

We just have to improve the process of requesting, approving, and revoking access, so that at all times you only have access to the data that you need, whether it's to build a model, to run a model, to investigate data quality issues, or why the model is broken and nothing else. But just getting, requesting, and accessing the model should be easy and automated where possible, so that at all times it's just limited to what you really need.

# DPO and CISO roles

37:19

Alexey

**Yeah. Well, in defense of the company where I worked, we actually had data protection officers. I was not always given admin rights. Maybe to speak about this data protection, I remember having a discussion with that person when I needed to access some sensitive information like emails. All of a sudden, I got to know this person (data protection officer). He was pretty interested in why I needed this data and I needed to explain to him why. Do you see this role often – this data protection officer role?**

37:57

Bart

I see them as an important stakeholder. I see companies that I was talking to, where you see that the data access management projects actually started because DPO was concerned or the CISO is concerned. I mean, it's their responsibility. Actually, it's an organization's responsibility to keep their customer data private and secure. A good example is Optus in Australia, which has been breached a couple of months back. They lost 10% of their customer base after that breach. So customers are becoming more and more privacy-aware. As an organization, of course, you have to take that responsibility. But it's often driven by the DPO, who has to comply with privacy regulations. They have to tell things to the supervisor – they have to report on that. They have to tell the supervisor and the customer, “Hey, we only use the data for the purposes that we've agreed upon.”

And then the CISO has to make sure that there are no data breaches, no data leaks. They have to prevent unauthorized access. For operational systems, that's pretty straightforward. But when you look at your data warehouses, your data lakes, that's where they lose sleep, because you've got all that data, it's messy, it's a lot of data, and you have a lot of change. Then you also have to prevent unauthorized access. Data scientists, data analysts, data engineers have to do their job, so they need to have access. And on the other end, you have to comply with privacy and security regulations. That balance of doing your job, being productive with data, giving you a company that competitive edge with data, while keeping it secure and private – that's a very difficult battle and it creates friction, such as in the example of your DPO at your previous job being like, “What do you need to data for? Do you really need it for that purpose? Can you explain it to me?

39:58

Alexey

**He would ask these questions all day long.**

40:01

Bart

Yeah. And that sucks. It sucks for everybody, right? Because then you have to explain to them. Not all DPOs are very technical. They often work more with legal texts and with policies. So there's this miscommunication and it just turns into policing. It shouldn't be. It should be automated as much as possible. The collaboration between the DPO and the data team should be streamlined as much as possible. You need to have given the insights and just enable it rather than policing.

40:35

Alexey

**I was also curious... You actually mentioned two things. We talked about the DPO, but you also mentioned the CISO. Can you maybe tell us a bit about these roles? Who are these people usually? What's their job? Why do companies actually employ them? Why do they care about them?**

40:55

Bart

Yeah. They're typically of a certain size. When you're working with sensitive data, and as your company reaches a certain size, you will typically see the data privacy officer (DPO) and a chief information security officer (CISO). But what I've seen is that they've grown towards each other. They used to be two completely separate entities. Now, they collaborate more closely. That's because the DPO has to rely on the CISO. What does a DPO do? He or she makes sure that the data that you, as a customer, give to the organization, is only used for purposes you've agreed upon.

If I give you my data for billing (for sending invoices) and I say that you cannot use it for marketing, I hope that I don't get marketing emails from you. That would create a trust breech. The next one is not always that big, but you also have unauthorized access or use by third-parties – by hackers. And that's where the CISO comes in. The CISO makes sure that no unauthorized people can have access to your data. They keep everything secure. Your DPO makes sure that the data is being used in a way and in the context that you, as a customer, can rightly expect.

# How data access management works with a data mesh approach

42:20

Alexey

**Thanks. I noticed that we have three questions. I think it's about time we covered them. A question from a Iop. I hope I pronounced your name correctly. The question is, “How does one deal with access management in a data mesh setup with sensitive data? Usually data producers own it, but for sensitive data, even producers shouldn't access it.” Maybe we covered this a little bit, the first part, at least. Also, not everyone knows what data mesh is. Maybe you can introduce it in just a couple of sentences.**

42:54

Bart

Yeah. Great question, Iop, by the way. First off, data mesh is a new framework from ThoughtWorks, from Zhamak. It basically applies the best practices from DevOps and data governance, and it applies it to data. There are four principles. I don't know them by heart, I think it's: data as a product, domain thinking, self-service analytics, and federated competition governance. Not 100% certain.

43:22

Alexey

**I have a shameless plug. We also had Zhamak in an interview, so you can check it out. I think this may be enough information for those who don't know what it is. If you're curious to find out more, you can just go and check that one out.**

43:38

Bart

Indeed. Zhamak would do a much better job than I just did. But it's a great point. Like Iop said, often it's the data producers or the data owners that decide on who can have access to what. But there's this point of sensitive data, because that has to comply with privacy regulations, security standards, and support. The way that we see at Raito is that it should still be controlled by your central data governance team, who know the regulations, know the privacy standards, and it should be automated as much as possible. Then we're looking at column-level masking, or row-level filtering, where based on the central policies, these columns are automatically masked. A good example is the marketing department, which has datasets with customer data, and they say, based on their policies, “Well, John and Mary can access this dataset. I approve access.” But in that dataset, there's payment card information or home addresses, then there should be a central policy that says that these columns should always be masked. And that's something that should be automated as much as possible. By the way, we call this the Conway-Benson movement, where both teams collaborate on that topic. Great question.

# Avoiding the role explosion problem

44:55

Alexey

**Another question from the same person. We already spoke about role explosion, when we have too many roles, maybe when we forget to revoke access. The question is, “Do you have any recommendations for following the least privileged access principles without falling into the role explosion problem (when there's a role dedicated to each private dataset)?”**

45:21

Bart

There's a role for each dataset. Right. I see. Tough question. From what I've seen, role explosion typically comes when your roles diverge from how the business uses data. A couple of things there. We see work with organizations that have role inheritance. They have the data roles, which gives you low-level permissions on datasets (read/write permissions and so forth) and then they have the functional roles that correspond to your roles in the business that can then assume the data roles. That's one way. Another reason for role explosion is access requests.

People request access for building an ML model, for investigating issues – there's no role that matches that request perfectly, so what you typically do is just quickly create a new one, and then forget to revoke it. So I think regular reviews and alerts for unused roles will definitely help with that. I think having the role inheritance, where you have roles that match with business and roles that match with data, combined with regular reviews and alerts on unused rules – I think that should help.

# The importance of data governance integration in DataOps

46:42

Alexey

**Thank you. Make sense. Another question is, and I've been sort of guilty of this as well. “Oftentimes, data owners do not know if the data is sensitive or not, or maybe they accidentally make a mistake. How does one make sure that we have the right processes to catch PII data (personally identifiable information)?” I happened to be a data owner at some point, and produced data. I just didn't realize that some of this data is actually private or sensitive. Then somebody pointed it out after discovering it accidentally (there was no process for that). Then, of course, I went to our data catalog and marked this field as sensitive. So is there a process for that? How should we deal with these kinds of problems?**

47:33

Bart

Really, the best way to address this is by integrating governance in your DataOps process. Once, I was talking to this guy who used to be in a data engineering team at Woolworths in Australia with 200 people. He said, “When I started integrating data governance in my DataOps, the effort became 1% of what it used to be.” As you're creating the data product, you add the tags, and then based on those tags, you can have an automated decision process that masks sensitive data. Ideally, the data is tagged – that's the ideal scenario. Then data is automatically masked based on those tags. Of course, you always need regular reviews. I think having good insights in who has access to what, what the usage patterns are and showing that in an easy to consume way, will also be very important to catch issues like that.

48:38

Alexey

**But I guess the question was also about, “If I accidentally mark data as not sensitive, even though it is sensitive, is there a way to automatically detect that?” Or is it that we just need to work on processes? Maybe we just need to have regular reviews?**

48:57

Bart

So about the tagging itself. Right. There's now this concept of 'active metadata' where the metadata is actively matched. You have a lot of automation there. Models to help you with the tagging. Not really my field of expertise, but I think there are several solutions out there that help with automated data tagging, which can help you with that.

49:26

Alexey

**With emails, addresses, phone numbers, IP addresses, geographical coordinates, etc. it's kind of obvious. They follow more or less the same pattern, so there are probably automatic ways to detect that.**

49:41

Bart

Yeah. These models that are out there, they can indeed use regex patterns to detect sensitive data, but they can also use the context of the data itself and look at the actual values and add tags in that way. There's a whole field that has been developing in that space over the past couple of years. Some of them are even open source. There are definitely some solutions out there to help with that.

# Terraform as a stepping stone to data governance

50:08

Alexey

**Now I want to talk about implementing everything we talked about. What I typically see in companies is tools like Terraform, or even before Terraform – we have all these clouds, for example, AWS cloud. In AWS, we have this thing called IM, which is...**

50:27

Bart

Identity and Access Management.

50:31

Alexey

**Right. Then what typically happens is, we use this infrastructure, which is called tools such as Terraform, or CloudFormation – all that. There is this infrastructure team that manages this code base. And if I need access to something, I need to poke these people and say, “Hey, can you please?” I have a role (or my team has a role) I want to have access to this S3 bucket, so I need to poke them first. Then they say, “Okay, we're too busy. Just create a pull request to this thing.”**

**So then I go and make sure that I can actually run this thing on my computer. Then I may add this to the Terraform file, then I create a pull request, and they take some time to review this pull request – somebody needs to approve it. So then somebody approves it, then somebody needs to apply it. Then maybe one month later I finally have access, if everything goes smoothly. I guess that's not the most ideal scenario, but from my experience, I see that often this is how it starts. Right?**

51:34

Bart

That's correct. That's indeed how it starts. I think it's a good way to get started – with Terraform. Because for most data teams, like I said a bit earlier, it's all about proving the value first. So you're really focused on getting those first data products out, proving the value of data analytics, getting more data consumers on there, more data products on the platform as well. Then that initial Terraform script that you used to manage access requests and process to request access to Slack or email – they start breaking. It doesn't scale.

That's because 1) It's too technical. For data owners to take ownership, they can't go in the Terraform code and start managing access, so it keeps on being with the core data team. So it's too technical. If by the time you have a DPO or CISO, there is no way of reporting – there's no insight, there's no visibility, and there's hardly any automation. 2) You cannot automatically approve or revoke access. 3) An issue I heard at a tech scaleup was that they had state drift in their access controls. They had access defined in Terraform and it was different from what actually was in Snowflake, because other engineers were just directly updating permissions in Snowflake. So I think it's a great way to get started. It just gets you started. You don't have to worry about access management. That is covered.

As of a certain scale, when you start looking at more vendors to help you manage permissions, the way is – we just build on top of that. I really believe in access as code as much as possible – pushing governance left. My vision is, “Let's just keep defining data products as code and the roles that can access your data products as code. Then everything like access requests, the management, the reporting – let's do that in a dedicated tool like Raito, where everybody can go to request access and your data engineers are just taken out of the process and can focus on cool stuff instead of processing access requests.”

# How Raito can help an organization with data governance

53:50

Alexey

**I see that we don't have a lot of time, but I still want to talk about Raito. I'm just wondering – engineers being engineers (and I'm one of them too) why do I need a vendor if I can just implement this thing myself? It's so much fun, right? [chuckles]**

54:06

Bart

Yeah. Building is indeed fun. Maintaining sucks. You have to update your connectors all the time. It becomes costly. But then also, just for your employer, security is important. I've seen several cases where the engineer that built it left and nobody maintained it. He left with all the knowledge. They get huge key person risk, and on something as important as data access management, it's not so ideal.

54:39

Alexey

**So how do you solve this problem? What do you actually do?**

54:42

Bart

The biggest challenge with implementing data access management is change. People have to change their behavior, change their process, and maybe even change tools. The way that we do it is by limiting the change that it needs, gradually improving data access management. And the way to do that is by letting you start from your as-is situation. There's no 'big bang' change. You don't have to disrupt everything. We come in and we have a free version of the product. You integrate that with your data warehouse or your data lake and from the first day, you instantly see who has access to what and what the usage is.

From that place (from your as-is) you can gradually implement or improve your access controls through collaboration, insights, and automation. That is like the opposite of what other vendors do, which is, “Okay, firstly, find your policies, your metadata, and then push down all that change.” And it breaks a lot. There's a lot of change all at once. That big bang is something we don't believe in. We believe in gradual implementation.

# Open source data governance tools

55:56

Alexey

**Are there open source alternatives – if somebody is not ready, or somebody just wants to make sure they see all the code. Are there alternatives for that?**

56:08

Bart

We have a Raito CLI that is open source, so you can manage access code. Now, we have our first contributor, who is building out the report capabilities. We're super happy about that. In terms of other open source, what we see all the time is Terraform.

56:30

Alexey

**But Terraform has all these issues that we talked about, right?**

56:33

Bart

Yeah, indeed.

56:36

Alexey

**Well, since you have an open source thing – we have a thing called Open Source Spotlight, where we invite open source authors to demo the tools. You're more than welcome to them or Raito CLI. We can organize something. I think that's all we have time for today. So thanks, Bart, for joining us today and for sharing all this information with us. And thanks, everyone, for joining in, for listening, for asking questions. I guess that's all. Thanks, Bart.**

57:07

Bart

Thanks, Alexey. Thanks for having me.

57:09

Alexey

**Okay. Well, see you soon. I don't know if it will be you or somebody else from your team presenting Raito CLI, but I'm looking forward to that.**

57:18

Bart

Yeah, that will be our Dev role.

57:21

Alexey

**Okay, that works. Thank you. Bye.**