1:56

Alexey

**This week, we will talk about data engineering and fraud detection. We have a very special guest today, Angela. Angela is a data engineer working at Sam's Club with their fraud team. She has four years of experience as a data engineer, and she's currently specializing in machine learning for fraud prevention. She works on designing and maintaining the data for a machine learning system that identifies fraudulent transactions. Welcome to the show!**

2:30

Angela

Yeah, glad to be here.

2:33

Alexey

**The questions for today's interview were prepared, as always, by Johanna Bayer. Thanks a lot, Johanna, for your help. Let's start.**

# Angela's background

2:41

Alexey

**Before we go into our main topic of data engineering and fraud prevention, let's start with your background. Can you tell us about your career journey so far?**

2:50

Angela

Yeah, for sure. I've been a data engineer for about four years now. I started out in Sephora as a data engineer, where I worked more with e-commerce and marketing tools but then I later… [cross-talk]

3:00

Alexey

**That's the store for selling cosmetics, right?**

3:04

Angela

Yes, exactly. Luxury, cosmetics – those types of items, yeah. Then I moved over to Sam's Club about… I want to say about a year back or so. So it's going to be two years coming up next year. I've been working within the fraud space, which is definitely a bit different, but still with machine learning. You can still utilize it, it's just in a different aspect.

3:33

Alexey

**So I know what Sephora is, because every time I go to a shopping mall, I can smell that Sephora is nearby. But what about Sam's Club? What do you do there? As a company, what does it do?**

3:45

Angela

Sam's Club is like Costco – it's a wholesale company. It's actually owned by Walmart. Walmart is general ecommerce, but Sam's Club would be considered to be a wholesale store, where you can buy things in bulk.

4:00

Alexey

**What's a wholesale store? I have no idea. [chuckles]**

4:03

Angela

It's just bulk items. You can buy cases of sodas, and if you have large events going on, it's really helpful. They also sell other things like electronics and stuff like that. But I think it's considered wholesale because you're able to buy things in bulk. A lot of restaurants and other businesses would work with Sam's Club.

4:30

Alexey

**Okay, let's say I want to buy a can of soda (Coke) I can only buy a whole pack?**

4:40

Angela

Yeah, exactly. It has to be more of a bulk item rather than a single-use can. Yeah.

4:48

Alexey

**So instead of one chocolate bar, it's like 100, right?**

4:52

Angela

Yes, exactly. It would be at least a case.

4:57

Alexey

**Yeah, I think I saw stores like that. They all sort of require you to be members. Maybe it's your competitor – Metro. At least… I don't know if they're a competitor, but this is what I see in Germany and other countries in Europe. The stores are called Metro. Usually, when you enter the store, they ask you for some sort of paper proving you are a company or whatever. It's not like you're an individual who wants to buy chocolate for cheaper – they want you to be a company, as their client.**

5:35

Angela

In this case, with Sam's Club, we allow anybody to be a member. It's really helpful for homes that have a lot of people in it, or maybe have a lot of roommates. So those are different types of use cases where [going to] a wholesale store would make sense. I actually do purchase some items from Sam's Club – usually, protein bars or protein drinks, or even specific types of sodas that I can't find anywhere else that are far more expensive as a single-use [can].

6:10

Alexey

**Yeah, I would also buy protein bars in bulk. [chuckles] I buy a lot of them anyway. [chuckles]**

6:16

Angela

Exactly! Yes. It's cheaper that way because they're far more expensive [as single items].

# Angela's role at Sam's Club

6:22

Alexey

**So, as a data engineer, what do you do there? You mentioned fraud.**

6:28

Angela

Yes, I specifically work within the fraud realm. A lot of the time, as within any company, as a data engineer, I will be working on maintaining their data pipelines, their different types of jobs, such as batch, or real-time jobs, trying to make sure that, if the business requires a specific field or specific dataset – trying to design that data set for them. If there's specific analysis involved with our machine learning models, making sure that we have dashboards and data feeding into those dashboards, where we're able to easily understand our metrics, from both the ML level and the business level.

7:18

Alexey

**I'm not sure into how much detail you can go, but I was just wondering, in the case of these wholesale stores, what kind of fraud is there? Does somebody pay with stolen credit cards? Things like that?**

7:38

Angela

Yes, exactly. That's a pretty good use case. It's very similar to credit card companies where we're trying to understand if someone steals your identity, or someone steals your credit cards, and they start purchasing from Sam's Club, how can we prevent that fraud? In addition, there are specific policies that exist within different companies, such as different return policies, or different shopping policies, where we might notice a trend of member abuse and we might want to block that.

8:13

Alexey

**For example, if I buy a pack of protein bars, eat half of them, and try to return the rest, then it would be fraud, right?**

8:21

Angela

Yeah, exactly.

8:24

Alexey

**Okay. So what do you do there? Maybe you can tell us in more detail – what kind of… I guess, when it comes to fraud detection, you want to react as quickly as possible, right? Your pipelines need to be real-time, right? Maybe you can tell us a little bit more about what exactly happens when you implement some of these use cases?**

8:51

Angela

Yeah, for sure. We have a batch job that runs on a daily basis that will actually make calculations for our feature engineering. We have the batch jobs for our feature engineering jobs and that data would then be used during our real live system, where if a member is making a purchase, there will be a call to our fraud system that will make a decision on whether or not, based off of the criteria of that transaction, if it is fraud or not. So depending on what we end up sending, that portion is definitely live – us sending that request. If it is fraud, then we will block them from making that type of transaction. If it's not fraud, then they'll be able to make that transaction.

# The usefulness of knowing ML as a data engineer

9:48

Alexey

**I'm just wondering – in your case, you work with machine learning. Do you need to know a lot of machine learning yourself? Or are you more focused on… For you, it's more like, “This is the use case and I focus on the data pipeline. I know that there is some machine learning happening – there are data scientists who deal with that – but I mostly work on making sure that the data is always there for these data scientists to consume.”**

10:14

Angela

I do also work on some portion of getting the metrics evaluations during the real-time process. We do end up sending our request and our recommendations out. Then after we do that, we'll want to make a calculation based on how the model is currently performing. So anything that has to do with more of an operational side, or more of the deployment and handling after the deployment – that's on my team. My team consists of both data engineers and machine learning engineers. We deal a lot with machine learning operations as well as data pipelines and creating dashboards for our ML metrics. Our data scientists do tend to focus more on the model development itself.

11:14

Alexey

**So you're a part of the engineering team and then there is also the data science team.**

11:19

Angela

Yes. Yeah.

11:22

Alexey

**And by “your team,” you mean you are leading this team? Or do you work in this team as a data engineer?**

11:29

Angela

I work on this team as a data engineer – my lead is a machine learning engineer. I work within that team itself.

11:43

Alexey

**How much machine learning do you need to know for this? Do you need to know any [at all]?**

11:50

Angela

Having a good background in machine learning for this role is actually really helpful. I think it makes it helpful to understand how to do certain analyses, but it isn't necessarily required. I work on the side of both analysis and maintaining the pipelines themselves, but I have some members who will work on setting up the pipelines, working with Kafka more so, and working with those types of access and databases. I think with my machine learning background, I do end up working on analysis and ad hoc queries. The main thing that I work with is considered the network model, so a lot of the stuff that I end up doing is trying to analyze the performance of that network model and other tasks around that.

# Angela's career path

12:48

Alexey

**Maybe I forgot, but I don't think you mentioned your background. So what is your background? I know that you also do something with NLP, right? I assume your background actually does involve machine learning. You know it, but at work, you are more focused on the engineering part, right?**

13:11

Angela

Yes, I actually majored in cognitive science with an emphasis on human-computer interaction and artificial intelligence, which, I want to say, is more user experience-focused, as well as more AI-focused. But I had a minor in computer science. During my undergrad, I focused a lot more on working with big data and understanding the impact of that. I did end up majoring in natural language processing. I think for a lot of NLP, currently, machine learning has been one of the more recent approaches that I find to be very useful – like the transformer-based models and more recently, ChatGPT. So I've gotten some experience with that.

14:02

Alexey

**Just curious, if you focused on NLP and machine learning, how and why did you end up doing data engineering?**

14:14

Angela

Yeah. When I was at Sephora, that was one of the roles that I ended up working towards. I think a big reason why I tend to work with data engineering is because I really love the importance of data within the ecosystem of NLP and ML. A lot of times people don't really understand that data really is the root of these processes. Even ChatGPT – in order for it to be able to be ChatGPT, it should have been trained on a bunch of data, there must have been pipelines that were scraping the data, putting that data into its training loop, and being able to maintain it on a daily basis as well. They are probably also getting requests on a daily basis, where they're having to consume that data and then be able to analyze the performance of their models, and then use it for their new methodology of reinforcement learning using humans' feedback. So data is really a part of all the ecosystem.

That's really a big reason why I like to focus on data engineering – because I feel like I don't necessarily have to work on the model, but it lets me understand what's going on with the models themselves. Maintaining the data, in my opinion, is one of the most critical components that you can have. If there are any failures, if there are any quality issues, your system will fail, and your system will have issues, and your models won't really work. So you might think your models are working perfectly fine but if your data is not there, they're they're going to fail.

15:54

Alexey

**Did you join Sephora as a data engineer, or did you first join as a data scientist and then transitioned?**

16:02

Angela

It's interesting, I actually joined as an intern – as a continuous process improvement intern. So I worked with... [cross-talk]

16:08

Alexey

**Continuous process what, sorry?**

16:11

Angela

Continuous process improvement intern.

16:14

Alexey

**What does that mean? [chuckles]**

16:15

Angela

Yeah, I thought the same thing when I joined. They told me that I was going to be identifying how to streamline processes. So I learned more about diagramming and system design. I used a combination of system design to identify pain points within whatever system they were [using].

16:40

Alexey

**Process modeling?**

16:42

Angela

Yes, yes.

16:43

Alexey

**I remember seeing something like BMPN – business process modeling notation or something like that. Did you need to do these things?**

16:55

Angela

Yes, I did end up having to learn about... I would talk to people about their business process, how they would work with a specific tool, and the issues that they would have with that tooling, and based on that, I came up with different suggestions or more of a holistic view of what was going on.

# Transitioning from data analyst to data engineer/system designer

17:18

Alexey

**So it's a bit higher-level work than being a software engineer because you need to analyze the whole system and then document it and then understand how it works and propose improvements. Okay, this is what you joined to do, but how did you end up doing data engineering? Because it sounds even less… not relevant, but… It's like data science and what you were studying is closer to data engineering than that, probably. But maybe it's just my ignorance.**

17:57

Angela

Yeah – I then worked as a data analyst for the company and then they were hiring for data engineering internally, so I was then able to talk to that manager and he was willing to hire me on as a data engineer because I wanted to work more with a tech stack. As an analyst, I was working more with things like Tableau, I was trying to build out how we could consolidate data sets and databases together in order to get different metrics for the team that I was working with, which was basically like data engineering in a way. That's when I really wanted to make that move.

Honestly, I would say, that the system design, has been only useful when I have to document the type of pipelines that we've generated and created. So if we're noticing that data is going from point A to point B, creating that pipeline has been something that I've – and documenting how that pipeline looks – is where that skill has come in. So more on the architect side of things, if anything.

19:06

Alexey

**Okay, so this knowledge does help you be a better data engineer.**

19:15

Angela

Yeah, it's more of thinking about the overall big picture. I think it's also helpful when thinking about things like the timing of data engineering jobs – there are a lot of questions you have to ask when you're first working to create a data engineering job, like who your stakeholders are, because then that can affect the timings of when your data engineering job should occur, and etc, etc. So there are definitely things to consider when trying to work with it.

# Best practices for system design and data engineering

19:48

Alexey

**Okay. Well, what we talked about – the system design – I think this falls into the category of having good documentation that describes this. The other thing you mentioned – knowing who the stakeholders are – is also helpful. I'm wondering, what are the general best practices that you need to follow when you are a data engineer in order to make sure that the thing you're working on actually makes sense, is reliable, and working well? So apart from having good documentation and knowing who the stakeholders are, there are probably other best practices.**

20:30

Angela

Yeah, absolutely. I think that goes to the type of data that you're hosting, how much data you're passing through, whether it's… I think, when you're looking at database design, that's where good principles really come in, and understanding the type of data that you're working with because then that affects... For instance, if you're working with dynamic data, you're going to be wanting more of a document-based database.

If you're working with data that's more static, you're going to want something more relational. In addition, once you understand the type of data that you're working with as well, it'll help you figure out the type of model that you have – if you're working with network-based data, that's a network model, and you want to use something like neo4J and when you're working with relational, then you have to decide whether or not you want to use STAR schema or Snowflake schema, and whether or not that really applies in your case.

# Working with document databases

21:30

Alexey

**You mentioned this document database – if the data is more dynamic, then you need to document the database (if it's static/relational). I was wondering, what's your experience with document databases? Because I heard very mixed messages about things like Mongo, DynamoDB, or other more “document databases. What's your experience with them?**

21:57

Angela

I've only worked, currently, with relational databases – specifically, Cassandra, Redis, those types – and then I've worked a bit with network-based databases, like Elasticsearch and those types of datasets.

22:17

Alexey

**So Elasticsearch is more like a network-based [database]?**

22:20

Angela

Yeah, exactly.

22:22

Alexey

**Oh. Okay. I always thought of Elasticsearch as more of a document database, because you index a document.**

22:30

Angela

Yeah, I think you can, but I think the way that we've used it… Yeah, we've used it with indexing as well. I think since I've worked with it with network-based data. But yes, it's indexed and it has a specific structure there. So I think that's why I kind of think about it more as a network, even though we probably just adjusted our network data to be more document-based.

22:56

Alexey

**Can you give an example of how exactly you put a network (and what kind of network) you would put in Elasticsearch?**

23:04

Angela

So this would be like Wikidata.

23:08

Alexey

**Wikidata?**

23:09

Angela

Yeah. For the Amazon Alexa Prize competition – I've worked on that for my Master's program – and even more recently, just for mentoring a couple of students. They work with Elasticsearch to be able to store their entity-based data. For instance, let's say you want to store information about books, like the author of a book, then you also want information about the year that book was created – we would want to store that book name as a document, and then it's relations within that database as well. So when the system's actually running, we're able to easily retrieve that information using different SPARQL queries.

24:04

Alexey

**There was a lot of information… I think I've heard SPARQL before, but maybe for those who did not… First, you mentioned Wikidata, right? What is Wikidata?**

24:19

Angela

Wikidata is actually built on top of Wikipedia. Wikidata has information regarding different entities. It's similar to Wikipedia, where you'll have editors of this network graph. To be able to gather this information, you need to be able to use this query-based language called SPARQL. It's very similar to SQL, but it has its own specific way of relating relations and entities together, and even getting inverse relations from the graph itself. So it's just a different way of querying data, based on the structure that it has.

# Working with network-based databases

25:03

Alexey

**From what I remember about Wikidata – you have Wikipedia, and the data on Wikipedia is not structured. Most of the time, it's just free text somebody put – an article about Game of Thrones, the book, for example. In the book, usually, people just write some information, and there is this table that is kind of structured, but it's free text – who the author is, when the book was written, and so on. There is a bunch of information there.**

**Wikidata is an attempt to get this information from unstructured (or semi-structured) data and put it in a machine-processable format such that you then can just run a query [on it]. Can you give an example of a query that you can run? For example, with books.**

26:05

Angela

Just select the book name, and then you would just use the relation of “what is book” and then that's where you'll do “?book” and then the relationship for that. Then you'll have the actual book name here. Because if not, then you'll just get the entity node. So that's what you want to grab. So you'll just push back “book name”.

26:39

Alexey

**I'm just wondering what kind of queries in addition… For example, I have this Game of Thrones book and I want to find all the other books by the same author, and then I can easily express that. Or books that were published in the same year.**

26:56

Angela

Yeah, so you'd want to do some sort of filtering statement, where you'd want to then filter by year, and have that specific year there or that specific name. That's exactly what we ended up trying to do – during the conversation, you might want to bring up other books related to Game of Thrones like, “Did you read the following sequel (the next book)?” Or “Have you read this book, which was [published] in the same year?” Then you would want to construct your queries to be able to filter these books based on the year and then send those titles back. Then you would just grab one, to then talk about it with the user. Then, if you want to pick it even further, you can do filtering on top of the genre itself. Let's say you want year and genre, to make it a bit more relevant, you would just alter your SQL query to be able to filter for that information and then give you that list of books to then recommend.

28:00

Alexey

**Okay. As I understood, in this example, we have a bunch of nodes. The nodes are different books (or just objects in general). Then we know that these books have links between each other and the link could be something like the author, right? For example, there is the entity, “George Martin,” and there is the entity “book,” and then we know that the relation between these two entities (these two nodes) is that George Martin is the author of this book. Then there is another entity, “year of publishing” and we know that “Okay, this book was published in this year,” and so on.**

**Then we have this network, and in this network, we can easily answer questions like the one you mentioned. In your case, you mentioned that you also currently work with network data at your work, right? So is this something similar? I don't know. Maybe you actually do that – I don't know into how much detail you can go there. You probably don't want to reveal how your flood prevention system works, but maybe you can give some details without going too deep into what kind of network you have there.**

29:15

Angela

That would just be relating the members to other members, or relating transactions to other transactions using specific correlations that can be similar to one another. Let's say we want to get all the transactions and we want to see if maybe there's a trend with the products involved. Maybe we want to get all the similar transactions with similar products and then based on that, if there are specific transaction networks where we notice that these transaction networks and these products have been related to certain fraud cases, then we might want to maybe use this as an additional vector, or use this as like additional information.

30:16

Alexey

**Do understand correctly? We have a network – in this network, we have transactions, we have products, we have customers – all these things are nodes in this network. Then we know, “Okay, there is this transaction with this number that involves this customer – it involves these sorts of products.” Then we have this network. We see all the customers, we see all the transactions. We know, in some cases, what the fraud cases are. We know, “Okay, this transaction was fraudulent.”**

**What that gives us is access to other transactions that are similar and maybe they're fraudulent, too. Maybe this customer (or customers) that are similar to this customer bought this product – maybe these transactions are also fraudulent. Then we can send them to our machine learning model for checking. Is this how it works?**

31:17

Angela

Yeah. It could be either an additional... There are many – I'm trying to be very vague, and nonspecific. Because you could just add it as a layer based on this type of network criteria that we would send to the model. Or you could even use it as an additional feature. Or you can use it during your analysis and if there's a very specific trend going on, where specific combinations of member and transaction networks, maybe you'll want to block certain transactions from happening based on the networks that are created.

# Detecting fraud with a network-based database

31:56

Alexey

**How do you actually know if a transaction is fraudulent? Is there some information coming from a bank that says, “Okay, this is a stolen credit card”?**

32:04

Angela

Yeah, that's a great question. Usually, we do get that information about whether or not this is a stolen credit card, or whether or not this is a fraudulent transaction. It's not something that we recognize immediately but that is something that we ended up looking towards.

32:22

Alexey

**I imagine that a credit card gets stolen, then somebody goes and buys a lot of Coke and drinks all of them, but then the owner of the car discovers that the card is missing in only a few days. Then they report that and by the time you have this information, the Coke is already gone.**

32:45

Angela

Yeah, exactly. So that's where our model failed and then we would have to identify “Why was our model failing in this case? What can we do next that can make this even better?”

32:58

Alexey

**But ideally, what you want to have is a situation where somebody is paying with a card and then you already have somebody from security say “Hey, hold on. Can you wait a little bit? We want to check your transaction?”**

33:11

Angela

Yes, exactly. Yes. [cross-talk]

33:13

Alexey

**So before they walk out.**

33:15

Angela

Yeah. For instance, if the card that they're using doesn't necessarily match their membership information or there are certain things about it that might flag something, then we might have specific protocols in mind that would get flagged.

33:34

Alexey

**Well, I guess here, since it has to be real-time system – you have to be really fast – how do you design a system with these requirements in mind, such that when somebody tries to pay with a stolen credit card, the security guard can get notified immediately? How should the architecture look for this sort of system?**

33:58

Angela

I would assume that it wouldn't necessarily start with a machine learning check, to be honest. I would assume that there would be specific roles, which I guess could be machine learning if it's role-based. But we could start with specific roles in mind, looking at those checks, and then you would have to work with your front-end team to be able to actually show some sort of signal to your cashier, which can then be triggered to your security worker, or which will then flag a sort of wave of different protocols. So I'm assuming it would have to be multiple systems trying to communicate with each other based on that one call.

34:46

Alexey

**I guess there's a system that tracks all the transactions and puts them somewhere, right? You need to have a reliable way of capturing this data and putting it in such a way that it's easy to retrieve and store.**

35:02

Angela

Yeah. That's why we have our batch jobs that are run on a daily basis. These batch jobs allow us to – we've already made the calculations themselves, and then we're able to bring forth that information. Then, in addition, within our system itself, we'll make real-time calculations based on information within the payload. That will happen almost instantaneously.

# Selecting the database type to work with

35:33

Alexey

**Well, I guess there's a lot of stuff that we can go into and a lot of details. I don't know if it makes sense to talk about it in a podcast format – it's probably very difficult to understand all that. But I'm wondering – in this case, you use a network database, but sometimes you can just use a relational database, right? Maybe in some cases, it will work even faster, instead of traversing all these nodes. You know in advance what kind of queries you will use, so you build your relational database in such a way that it's very easy and fast to answer these queries.**

**I'm wondering, how do you make a decision on which database to use for a specific use case? Should you go with key value store? Should you go with document store? Should it go with a traditional relational database? Do you have a rule of thumb or maybe some criteria you use for evaluating this decision?**

36:35

Angela

Yeah, I think that's where your type of data comes in. If you notice that it's not going to ever change, your schema will probably remain stagnant, then relational will probably be the better case to go with. If you're noticing that dynamic – that's when you have to actually think about the document. I've never worked with dynamic data. But that is, from my experience, the criteria for that. Then in terms of more of the key value store, it's just a matter of what the use case that you're working with is and whether it actually makes sense for that type of data point that you're trying to capture. Because if in the end, you're just trying to put everything into tables, I can't see why you can't use a relational database.

But if you need a specific type of analysis, or a specific type of structure for your data – I think that also depends on the structure of your data as well, actually, because we've mentioned like unstructured versus structured. With structured data, that's perfect for a relational database, especially if it's static. That's a specific criterion for those types of databases. But with unstructured data, you can always find a way to take unstructured data and turn it into structured data, which is possible, such as taking it from... maybe you have like a JSON. But if there's an additional layer on top of it, that's where I think you have to rethink the type of database that you're working with, or maybe get a bit more creative there.

# Neo4j vs Postgres

38:11

Alexey

**What's your experience with working with Neo4j? I heard things like, “Okay, it's a good database, but there are not so many people who have a lot of experience working with this database.” Sometimes what happens is when you hit some problem with this database, it's a lot more difficult to find a workaround. Did you have to experience something like that? Or for your use case, it was actually a good decision to use Neo4j, instead of, let's say, Postgres?**

38:45

Angela

I think for our use case, it makes more sense to use Neo4j, only because of the type of visualizations that you can be able to do with this. I would say that you could definitely have some problems in terms of understanding how to query your data correctly. But I think if you're working with data where it's pretty understandable and pretty easy to work with and your relations aren't necessarily too complex, it could be really helpful in trying to understand the type of networks that are being formed.

39:19

Alexey

**So this visual aspect is… because you can explore your network (your graph) by clicking. [Angela agrees] This is probably useful for… In the company where I previously worked, we also had an anti-fraud department. In addition to machine learning engineers, data scientists, analysts and so on, who would work on solutions, there were also anti-fraud specialists who were not super technical – they were more like domain experts – and they were one of the stakeholders (the users) of this tool.**

**Sometimes they would need to decide, “Okay, does it look suspicious or not?” Just by looking at some data, they need to make a decision like, “Should we ban this user? Or should we set them free?” For them, it was important to actually be able to open a graph like that and traverse it to see, “What are the other connected users?” In your case, it's something similar.**

40:25

Angela

Yeah, that's exactly it. It's really a great tool for end-users, really, and trying to visualize the network graph. We can always give them tables that can easily try to give them sort of that snippet of information. But there's a lot more complexity on their side and a lot more time that they end up spending having to look at this data in a more tabular format.

# The importance of having software engineering knowledge in data engineering

40:50

Alexey

**Yeah, I imagine. I noticed that there are a few questions from the listeners, so I think maybe we should cover them. The first question is, “How much software engineering is required when you work as a data engineer?”**

41:09

Angela

I think there are good software engineering principles that you should be able to hold when working as a data engineer. Because we do end up working a lot with either Scala or Spark – Spark is probably one of the biggest libraries that I've worked with. I specifically work with PySpark. But in my previous company, we worked entirely with just Scala Spark. So you should be able to at least work with these different languages and be able to construct your code in a way where it makes sense. I think, generally, as a data engineer, you should also think about the different types of testing that you might want to complete. I think software engineers really know how to create good unit tests to really test the system as a whole.

But we would have to think about it from the data engineering perspective like, “How can we add in data quality checks within our pipeline on top of the dataset that you're creating itself?” So if you want to add in null checks, if you want to add in checks for the specific date types, or specific types within your schema to make sure that there's no failure there – that's where having a really good ability of using test and looking for different libraries is really helpful. So there are definitely software engineering principles and practices that you'd probably want to utilize with data engineering.

42:43

Alexey

**So what you said is that there is quite a lot of software engineering knowledge that you need to have – all these programming languages, testing, best practices – but there are other things in addition to that, which are data quality checks, null checks, etc. Would you say a data engineer is something like a specialized software engineer – one who specializes in data?**

43:08

Angela

Yeah, I think that's a great way of putting it. As a software engineer, usually people think about the application or the system like, “How could it fail?” So you then have to put on a different hat and instead of thinking about the application, you think about the data itself. So yeah, that's a that's a perfect way of putting it.

# Data quality check tooling

43:28

Alexey

**Just curious, what kind of tools do you use for these data quality checks?**

43:33

Angela

You can use something like Great Expectations. You can also build your own types of unit tests, or your own framework if you'd wanted to. Great Expectations is one of the ones that always comes to mind for me, but I'm pretty sure there are other third-party companies or toolings that have that in place. Even if you look at other data profiling tools and stuff like that, they might also have a layer of data quality. Even cloud services like Google BigQuery, they have data quality checks integrated within their platform itself. So you don't necessarily need a specific library or to add in those tests, but you need to identify how you can start adding in those tests into your pipeline.

44:24

Alexey

**So as a data engineer, you need to be aware that these things are important, and you need to think about them when you design your data pipeline so that you don't forget to include the data quality checks.**

44:40

Angela

Yeah, exactly.

# The greatest challenges in data engineering

44:41

Alexey

**What are the most challenging tasks in data engineering practice?**

44:51

Angela

That's a really good question. I think one of the most challenging tasks is when 1) Your job fails and you don't know what If your job is failing. Usually, as an entry-level data engineer, when you're first starting out, you're not really sure how a job specifically works – that's when you have to really look into your log files to understand whether it's the code. Because if it's the code, then it's probably a bug or a fix that you need to make. Is it a schema change that happened not because of you, but because of your incoming source of data? If your source has changed, I think that having that visibility is difficult sometimes because it can cause your job to fail. If, when calling your database, it's having issues with... maybe your database is failing because there are too many requests at a time and maybe it's processing too much data, so maybe you need to think about whether you need to make this an independent job.

So when your job fails, it really makes you have to think very creatively and understand why it's failing, what's going on, and whether it might be a bigger improvement that needs to be made during that time. I think that's one of the biggest challenges as a data engineer – we might make certain configurations based on the data load that we've tested with within the testing environment, but once we actually identify how many of the jobs that we actually have running concurrently, in parallel, and just these different aspects that could actually affect your system. Then you'd have to start reconfiguring, and re-optimizing your jobs – maybe it's an optimization within the code level. I think that's one of the biggest struggles as a data engineer – just trying to identify what the best solution is for the job that I'm working on.

# Debugging and finding the root cause of a failed job

46:56

Alexey

**You mentioned multiple reasons why a job can fail. The first reason is the code – there is a bug in your pipeline, in your PySpark job or in your script. Then there may be a schema change – there's upstream data that you consume and then the team that is producing this data changes something there. There could be an issue with that database – the database that you write to or the database that you read from. I think you're only scratching the surface here. There can be so many other reasons why it could fail, right? It could be an error in the code, as you mentioned. It could be an error in the data – maybe there is an erroneous record that goes through your pipeline. There is a null somewhere and you expect a number – all of a sudden, your entire job fails just because of that record.**

**There could be millions of reasons like that. Debugging (finding the root cause) could be a nightmare. I remember that. I understand why it's one of the most challenging things to do. So do you have an algorithm to figure out what the root cause is? How do you usually go about that?**

48:21

Angela

Usually, once you actually get accustomed to your jobs and the type of errors that you see, there tends to be a trend in how you solve them. I really think it comes with experience, because I think when you first start out as a data engineer, you're not really sure about these types of errors. But once you've actually worked with these jobs, you become familiar with them, then you start noticing, “Oh, it's probably a database issue,” or “Oh, it's probably something with my code.” There are also some times when your job doesn't fail – your job works perfectly fine, but the data is incorrect because of the upstream team. Maybe you're just not getting a column correctly anymore.

So it's very much a trial and error process, but once you understand the errors that you're working with and the data that you're working with, you can figure out the trends and the patterns pretty easily, I would say. I think if you're still struggling with it, I would suggest creating some documentation with common error types that you're seeing and then identifying what the solutions were to actually solve them. If you create really good documentation for your jobs, which you should, then you can create run books for how to actually solve them pretty easily so that not only you, but anyone within your support team, can be able to solve these issues pretty easily.

49:45

Alexey

**That sounds like a great piece of advice for any data engineer, and specifically, Junior data engineers who just started. When you have a task that is failing and your job is to figure out why – if they try to do what you just said, maybe you'll spend a few days trying to figure out what's happening, but if you document everything, and if you also document how to fix this problem next time, the entire team will be so glad.**

50:17

Angela

Yeah, exactly.

# What kinds of tools Angela uses on a daily basis

50:23

Alexey

**I think that there was a question that just disappeared. Let me try to remember what the question was. “What kind of tools do you use on a daily basis?”**

50:37

Angela

I specifically use GCP, and Databricks. I've worked with Cassandra and I work with TablePlus to be able to connect to that. I work with PySpark probably every day. I've worked a bit with Pandas, but I don't have to use that quite often within this current role. However, I know some people who love using Pandas as well, especially with the new PyArrow implementation change that they recently made with that. I primarily work with PySpark and GCP cloud services. I've worked with Dataproc and BigQuery.

51:23

Alexey

**Dataproc is a tool from GCP – it's like a Spark cluster?**

51:59

Angela

Yeah, it's kind of like… I know Azure has a similar cluster configuration, where you're able to run your job and your clusters that way. Dataproc is more of a “managing your clusters” tool. But I know they also have serverless management now within that tooling as well. That's something that people should look into as well because that's pretty powerful – having jobs that don't necessarily require servers.

52:04

Alexey

**So in this case, let's say we have a PySpark script (PySpark job) and we want to execute this job on a specific piece of data. Typically, the way it works is, that you need to have a Spark cluster where you can execute this job. Then you need to provision all these machines, you need to configure it, and Dataproc makes it easier. I didn't work a lot with ПCP – I worked with AWS. In AWS, there is a thing called ElasticMapReduce (EMR). Basically, by clicking a few buttons, you can get the Spark cluster and then you can execute.**

**The serverless thing that you mentioned is – instead of worrying about provisioning all these clusters and then managing them somehow, you just say, “Hey, Google. This is my PySpark job. Execute it somewhere. I don't care where.” And then it just executes stuff. That sounds pretty convenient.**

53:07

Angela

Yeah, exactly. Right. I think that's hopefully going to be the future of data engineering jobs because it seems like that would be probably the most ideal way to go.

53:18

Alexey

**I remember when debugging a Spark job, it's not always easy to, first of all, make sure that you have enough computers there (enough executors), then select the proper machines for the job, then make sure they have enough memory and all that – you could spend days tuning this. With serverless, I guess, it's just much easier. You mentioned that there is a PyArrow change in pandas. Do you know what that is? What kind of change was made?**

53:59

Angela

Yeah. With Pandas, it's now running with PyArrow, so it's supposed to be a lot more efficient when it's running. Before I think Pandas wasn't running in a distributed fashion, so it was running a lot slower for that reason. But with PyArrow, it provides the ability to run in a distributed fashion, which is quite helpful. Because if you're just running something on a single instance, and you're working with Big Data… I also think one of the biggest issues with Pandas is that it does have a cap with the data size that it works with. I'm not sure if PyArrow somehow handles that differently, but from my understanding, that's one of the drawbacks of using something like Pandas – you won't necessarily be able to load all of the data into a single frame. But PyArrow is more of allowing it to be run at a faster rate.

54:57

Alexey

**In a gist, it just became faster because of some internal changes. [Angela agrees] That's cool. You also mentioned another tool that you use regularly – Cassandra. I'm wondering, what are the use cases for Cassandra? Because for Cassandra, maybe my information is outdated, but you actually need to have a cluster, again – a cluster that you manage. Or maybe you get a managed cluster from somewhere. I'm wondering, “Okay, what is it good for? For what kind of use cases should I consider using Cassandra?**

55:35

Angela

Cassandra is like a relational database. So if you're working with structured data, if you're working with data where it's relational (it can be in tabular format) that's where it can be helpful. I believe it's also fault-tolerant and it's really good with scalability. But you're right, it does need the clusters – you do need a cluster to be able to run a Cassandra database.

56:04

Alexey

**So it's good for analytics or for transactional? I guess for transactional, you should use something else. But the main use case is more analytical stuff, right?**

56:16

Angela

Yeah, exactly.

# Working with external data sources

56:19

Alexey

**Okay. Well, another question is, “How much of a challenge is it to get data from external sources?”**

56:31

Angela

That's a really good question because I think it depends on what you mean by “external sources”. For us, it could be, even though our data is internal within the company, maybe you're working with an external team. So it's being able to identify what teams that you need to work with in order to get the different datasets within your database so then your analysts have that access as well. It can be difficult to find the person and then find good documentation on that data in terms of how to use it properly. I think that's probably a very difficult thing to look at.

Then, if you're working with businesses outside (third-party services), let's say, within my company, maybe we're working with a third-party fraud service, where they're also running a fraud check. We would want to be able to get that information into our system somehow and be able to read that in. Then there's that layer of getting the documentation from their team to be able to understand what's going on. I think one of the issues with that is just making sure that you're getting the data on a daily basis and nothing from their side changes. In general, I think working with any external teams, or any external third-party service, to get data – or even through something like an API call – one of the worries of a data engineer is, “Is their data going to change at any point without me knowing?” Or “Is this data not going to be able to be consumed after a period of time?”

So there are different talks that have to go on in terms of creating that type of data contract and understanding how their external teamwork works, as well as understanding, “Is this data going to be on a batch basis at a specific time? Is this something that's real-time or occurs every four or five hours?” And then understanding how that would impact your analysis and your jobs and all of that? Yeah, I think working with external teams does bring in some difficulties, definitely in terms of, “Who do I talk to? Where do I get this data? How often will I get it? Will this data ever change?”

59:03

Alexey

**So it is challenging, and you need to think about… you mentioned data contracts. You need to somehow understand how exactly the data should look, and you need to make sure that it doesn't change. The API that is already there does not change all of a sudden – what used to be a number does not become a string or whatever, right?**

59:26

Angela

Yeah, exactly.

59:29

Alexey

**I already hear the bells from a church nearby, which means that it's actually time we wrap up. There are a few more questions. Maybe my question to you is if it's okay that people reach out to you somehow and ask these questions.**

59:48

Angela

Yes. Feel free to reach out to me on LinkedIn, and I'll definitely answer from there.

# Angela's resource recommendations

59:56

Alexey

**Yeah, thank you. Well, maybe the last one before we wrap up. We talked about many things. We covered many topics. Are there good resources – it could be courses, books, or articles, that you recommend to our listeners if they want to find out more and learn more about these topics?**

60:17

Angela

My biggest suggestion – I love the O'Reilly books. I definitely suggest getting the data engineering best principles – so more of an overview book. I don't want to specifically say a specific book type in case people have their preferences. But I think getting something that's more data of engineering overview, getting something that's more of Designing Data-Intensive Applications – I think that's a specific O'Reilly book that I would definitely recommend. Then definitely something with PySpark, SQL, to be able to always look those types of questions up during your interviews.

61:02

Alexey

**I think I read this Designing Data-Intensive Applications book like two times. I think if I read it a third time, I'll still find a lot of information there.**

61:12

Angela

Yeah, absolutely. It's definitely a good one.

61:15

Alexey

**Okay, Angela, thanks a lot for joining us today. That was really great talking with you today. And thanks, everyone, for joining us today – for watching and asking questions. I guess that's it for today and have a great rest of your week!**