EPISODE 1299

[INTRODUCTION]

[00:00:00] JM: Hello and welcome to Software Engineering Daily. Today's guest is Chris Fregly. Chris Fregly has been on the show before. He's an excellent engineer. In the last episode that he came on, it was probably four years ago, and he had been doing machine learning at Netflix. He'd been doing data infrastructure and Netflix. And Chris's experience at Netflix led him to write this article about the PANCAKE STACK. Or I think it was also a presentation that he gave at several tech conferences. And the idea of the PANCAKE STACK was something like for like Presto, Apache Spark. Basically, PANCAKE STACK was an acronym. So if you count the number of letters in the words PANCAKE STACK, it's I think it's 12. And if you think about like a 12 letter acronym, that means there's 12 different technologies involved in the stack. And the whole impetus of that show was the idea that the PANCAKE STACK is this very long acronym that basically represents a gigantic field that is known as data engineering. And so data engineering four or five years ago was even more nascent than it is today. So data engineering is very immature today, but it's getting better.

And so what does data engineering look like in AWS? That's one of the questions of today's show. Although I think it's kind of a behind the scenes question, because we don't really talk about data engineering. I think what's important to note is that if you're on AWS, you have an opinionated data engineering stack, and that integrates well. So arguably, going with AWS, going all-in on AWS for data science, or for machine learning, is quite a compelling environment because it's so opinionated, because you have this full suite of data engineering tools that are integrated well. Now, once you have that, it makes the whole data cleaning and data architecture process much easier, which allows you to do a whole lot more with machine learning.

So if you have more time to spend on machine learning, you will get better results. Machine learning is this highly iterative process. It's kind of a dog fight. You're kind of wrestling with your data. And so if the data infrastructure gets out of your way, if the PANCAKE STACK gets out of your way, you have a huge advantage. And that's the compelling advantage. That's the compelling motivation that Chris Fregly is making this episode in his vision for *Data Science on*

AWS. *Data science on AWS* is a book that Chris Fregly wrote with Antje Barth. We're trying to have her on the show as well. It's a book that's like a comprehensive book about data science. It's an O'Reilly book, very reputable. And I'm really thrilled to talk to Chris Fregly once again. He's one of the foremost people in applied data engineering and applied machine learning. He does lots of stuff for AWS. He's sort of like a guy that moves between teams, and produces content, and produces new product ideas. And I think he's just kind of an internal entrepreneur within AWS. So I love talking to Chris. He's fantastic. I can't wait to have him on the show once again. And I think you're going to enjoy this episode.

[INTERVIEW]

[00:03:01] JM: Chris, welcome back to the show.

[00:03:03] CF: Yeah. Hey, Jeff, it's been a few years, huh?

[00:03:06] JM: It has been too long. You are one of the foremost authorities on making complex production data engineering systems work. And you've taken that to a, well, I guess, what most people call machine learning today. I think machine learning is, in many cases, kind of a higher level version of data engineering. Or actually how would you define that? How would you define – Where are the lines between machine learning and data engineering?

[00:03:39] CF: Yeah, so you certainly need data engineering to build your models. So you can't really do one without the other. I would say once you start to really apply math and start to derive your like insights, I think is really, because you can do your queries and sort of get kind of summary statistics. But when you really need to get some of the advanced analytics and more like predictive models, that's where you start to use machine learning.

[00:04:13] JM: If we think about the world of machine learning from the standpoint of AWS, AWS, at this point, has a finely honed sense of building products in a way that will be appealing to developers. What is the AWS perspective on how you build infrastructure as a service for machine learning?

[00:04:38] CF: Yeah, so as you know, like AWS is very much about building the like building blocks, right? And as we hear from customers more and more sort of higher level solutions, right? So they don't just want a feature store. They don't want these separate pieces. So in December 2020, for Reinvent 2020, there were lots of new features that came out. And really the entire pipeline came together in December 2020. So that was super exciting. And in fact, like we were writing the book, me and my coauthor, Antje Barth, were writing the book. And we actually had to pause because we knew about all the things coming out for Reinvent. We had to wait for all those, because we couldn't launch a book that didn't have full end-to-end pipeline.

[00:05:27] JM: If I think about the companies I know that do machine learning, I mean, a lot of them are running on AWS, but I would not consider them all-in on AWS from a data engineering perspective. Like if they're using Databricks, or if they're using Spark, they're probably using Databricks. If they're using Kafka, maybe they're using Confluent. If they've been doing machine learning for a while, they might be running their own TensorFlow stuff, or maybe they're running TensorFlow on GCP, since GCP probably has better TensorFlow support. Maybe you can correct me there. So it seems like the most machine learning companies that I have encountered, at least, please let me know if you have any counter examples, like the people that go all-in on AWS for machine learning and what they get. But I'd like to get your perspective on the heterogeneity. Like if I take a heterogeneous approach, I'm not I'm not AWS specific, how am I selecting tools? What is my process? Or how has my tool chain come together?

[00:06:27] CF: So one of the things – So coming into AWS, I had a startup before called PipelineAI, where we were very much like multi-cloud, we were hybrid cloud. We were all Kubernetes-based. We started to migrate to Kubeflow. What I realized in that startup was like pretty much all of my customers, even TensorFlow, and Pytorch, were all using AWS. And so I had started that company a few months before SageMaker was launched. And so I've been tracking SageMaker for those previous few years. This is five years ago now. But I think what you're getting at is we're not like necessarily known for having the complete like lifecycle and the ability to take your models from raw data and then deploy them. And that really did change in December. And I had caught wind of that. And that was one of the reasons I had joined AWS. And like pretty much all my customers from like PipelineAI we're using AWS. Even if you're using Databricks, by the way, you still have to choose your cloud provider, right? And so you're either using AWS and getting your data from S3. You can also choose like Google if you are in

like Google Cloud, or like Azure, of course. And same with like Confluent, as well, too, right? You still need to pick which cloud. If you're using like Elastic Cloud, there's still a choice of cloud. And typically you're choosing based on where your data is and trying to minimize transfer of data.

We actually just got off a call with the Snowflake folks. And they're a very keen partner with AWS. And one of the things we're talking about is how to make that transition smooth, right? How to get data in and out from like EMR, from SageMaker. And yeah, we've been working very, very closely with them. On the sort of open source side, we call it the ML on containers. So with like EKS, Amazon EKS, the elastic Kubernetes service, yes, we certainly support Kubeflow. I actually had a blog post last year that was highlighting. I think it was the Kubeflow, 1.1 release, and all the ways that like Amazon supports Kubeflow, and security, and yes, all that.

[00:08:51] JM: You mentioned Snowflake there. Snowflake really shown the potential of a focused vertical approach to infrastructure, basically taking a domain that AWS has validated and going all-in on it. And going all-in on it to such an extent that, from my point of view, they kind of have shaped the market. They've kind of said, "Hey, you can do this thing where you're like the PANCAKE STACK," which we explored a couple of years ago, where it's like, okay, the whole idea of the PANCAKE STACK, which was hilarious. Why I liked that episode so much was it's really not about what is in the PANCAKE STACK. It's the idea that this acronym is so long that it has seven letters. There are seven things in the PANCAKE STACK. And there's actually more things in the PANCAKE STACK. And because it's a stack of pancakes, you can have as many pancakes as you want. As many things in the stack as you want. Basically, the idea is data engineering is confusing and terrible. And Snowflake, from my point of view, I'd love to hear your narrative on this, Snowflake says, "No, no, no, no. You don't want any of that. You're going to put it all in the data warehouse, and then you're going to do whatever you want on top of that data warehouse."

[00:10:06] CF: Yeah, the Snowflake guys really lean into security governance, role-based security, column-based security. They even have differential privacy built-in. Yeah, yeah, that's a very interesting observation. And coming from Databricks, that was similar to Databricks as well too, where they previously were known as the Spark-like company. This is Databricks. And now they're more sort of Spark and AI data. And then with Snowflake, it was interesting to see how

much they rely on partners for machine learning. So Snowflake very much still is just a data company. And the ability to run SQL queries is really their sweet spot, and have really, really tight security by default. Everything is locked down.

And so I think what you're getting at Jeff is there's one approach where you can go build everything yourself, right? And like a lot of us start off that way. And yes, I remember working the data bricks booth many years ago, and this is right when Spark came out, or when like Databricks came out with their product. And people came up to me all the time at the booth and would say things like, “Why would I use Databricks? Like why would I pay for this when I can run my own Spark cluster?” And I used to think, “Yeah, that's right. That's right.”

And then, slowly, I would be speaking to these crowds like throughout the years and I'd ask, “Hey, who here uses Databricks?” And I would see more and more hands going up, right? Like, “Who here uses Snowflake?” More and more hands going up. And the whole time, I'm thinking, these are probably the same people that had come to me a couple years ago and we're like, “This is silly. I can run my own cluster.” But like very quickly, now you're running your own Spark cluster. Now you're running your own Kafka cluster. Now you're running your own Elastic Search cluster.

One thing I realized while building my startup was I was spending all of my time just trying to get GPUs to work with Kubernetes. Just trying to get my YAML straightened out and keep my services running and get my Elasticsearch cluster, get my Grafana and like Prometheus running. And I was always running out of disk. And I was always – So these days, you don't really have to do that. So maybe it's worth the extra 10 cents per like compute hour, or like whatever they charge these days, right? So it really comes down to, at like this point in my career, I really value my time. And I like to actually focus on the business problem. And I wasn't always like that. I always like the sort of platform challenges, systems challenges. But as you get more mature, like little more experienced, and see more business problems, and realize that's where the real value is. So maybe it is worth it.

[00:12:54] JM: You've written a book about machine learning on AWS. Give me the synopsis of the book.

[00:13:03] CF: Yeah. Really, the value prop for this particular book is an end-to-end sort of complete picture. This is why we had to wait for like December 2020, Reinvent 2020, to get SageMaker pipelines in there. To get all the pieces under one like roof, if you will. So they're all under SageMaker. It's getting data in and then transforming those features, and training a model. And of course, we chose natural language processing. Bert was just starting to become very popular when we were starting to write the book. We use a library called Hugging Face, Transformers, which is a close partner of like AWS. And we then deploy the model. We tune the model in production. And then we decided to add in streaming. So we use Kinesis. We use Kafka. We show many different pipeline options. So we show how to use the new managed Airflow, the new managed Kafka. And, yeah, all the different pipeline options and streaming options. So the book ended up being 500 pages. We were only supposed to write I think maybe 300. But yeah, working with the O'Reilly folks was really fun. They're a great group of people. Have a very, very large network, and really helped us promote the book and build a super high quality book.

[00:14:25] JM: Now that you have gone very deep at – I guess you're at Netflix, right? The was the PANCAKE STACK story, right?

[00:14:33] CF: Yeah. Yeah, we used everything there.

[00:14:36] JM: Right. Netflix, they really were pioneers in cloud-based data engineering in so many ways. Spinning up their own streaming systems, Mantis, I think. I think I did a show on Mantis a while ago. I don't even remember what that does. Lots of random things got built, because they didn't have this stuff productized, which today it is. You can go on AWS and you have all the pieces in place. You've hinted at this already, that basically, if you're building with well-designed abstractions that abstract away, running out of disk and like that kind of thing, the thing that you should basically never encounter as somebody, if you're not working at AWS, you should never encounter running out of disk if you're not totally cost sensitive. How does building in such an unconstrained way in the world of machine learning, or data science I should say? Your book is *Data Science on AWS*. Not machine learning on AWS. I misspoke. But if I'm building a data science application in modern AWS, not Netflix circa 2015, 2016, whenever you're building PANCAKE STACK, if we're talking five years later, how does the experience compare?

[00:15:57] CF: Yeah. And I still maintain good connections there. Things have like evolved quite a lot, specifically around pipelines. They have a project. I believe it's called Metaflow, Netflix Metaflow. And the data challenges are still there. Data engineering challenges are – So they still have, as far as I know anyway, forks of Presto. Different sort of optimized variants of Kafka, things like that, that work at the specific Netflix scale, right? And some of those patches just don't ever make it in. Either they don't really align with the master roadmap for these projects.

I think, to get back to your question, so there's always data challenges, and those don't change. And as far as I know, there're still quite a lot of different systems at play there, right? There's Presto, there's Spark, there's – Yes, all kinds of like Hive metastores that they're using. They use EMR quite a bit. And in terms of data science, because – So I specifically work with quite a lot of startups who just want to get their models out, right? And they don't necessarily want to have to string together EMR with a bunch of other services. All they want is to point SageMaker to their data and then derive their models from there and then serve those models. But then they start to get to the point where they are seeing quite a lot of data. And now they enter with the same data challenges. And then, yeah, they end up having to like integrate all these.

So one of the things that I've been – Yeah, I've been working with the engineering teams at like AWS is trying to merge more of EMR with SageMaker, or find a nice balance there. Right now there's sort of separate EMR, and then there's Sage maker, and we're starting to see some of those worlds come together. And there's all kinds of like cross collaboration, like PMs between these teams. Yeah, because you worked over at like AWS, so you know how the like PM org works. So they basically have folks on Sagemaker talking to the EMR team. They have folks on like EMR talking to SageMaker. So Netflix is a very special case. They have tons of data. They have fast moving data. They have their own requirements. And like, oftentimes, they're building patches directly with these project owners at like Confluent, and at like Databricks, and those kind of places.

[00:18:36] JM: So, in more detail, how much time am I saving? If I go from this PANCAKE STACK world – Because I remember doing shows early on when I started to do this podcast, and data engineering sounded miserable. It just sounded completely miserable. I remember talking to like Max Beauchemin from Airbnb. He had worked at Facebook. And I've talked to

people who did the early Hadoop stuff at Facebook. And just it sounds so miserable. And I'm trying to get a bead on how much of that has been cleaned up. How much are you still having to do? Like if you move to AWS, are you just moving the problem? Or is the time to market actually getting like significantly better?

[00:19:23] CF: Yeah. So at Reinvent 2020, there was a feature or service that was released called SageMaker Data Wrangler. And so Data Wrangler, really, to me is kind of the closest thing to that like drag and drop sort of thing, which not all of us buy into. I'm a little skeptical of like drag and drop. But what it's doing behind the scenes is very cool. It's actually launching little mini spark jobs to do these transformations. And so all you do is specify, “Here's my data in S3. It's of CSV type, or TSP, Parquet, whatever.” And then from there, I can make my transformations. I could save that. It's called a .flow file.

So from that standpoint, things become a little bit more communicable to your teammates. You're not passing around a whole bunch of bash scripts and Python files, R files, all kinds of weird things that like Max always talks about. I know Max. So Data Wrangler actually supports all of Spark. And once you realize that it's Spark behind the covers, and you're not managing those Spark clusters, these are serverless Spark clusters. So you don't even think about like scaling these things up. They just auto scale for you. And so you're paying just for the time that you're actually running those queries.

And from that standpoint, much, much easier to code, much, much easier – And the super cool thing about Data Wrangler that like not a lot of people like talk about is I can actually click export. End export, a big Python file, if I really want to have Python, or if I want to move this to another team, or if I want to move away from SageMaker Data Wrangler, if you want. And I can actually then commit that code into GitHub, and then I could use it outside of like SageMaker.

[00:21:17] JM: I have just finished a book of my own. And I know that reading a book is an arduous process. Can you take me through the various points in the torture chamber of book writing?

[00:21:36] CF: Yeah, which book did you write? Yes, I missed that.

[00:21:37] JM: Well, it's actually not even out yet. It comes out July 6th. It's called *Move Fast: How Facebook Builds Software*.

[00:21:44] CF: Excellent. Oh, awesome. Yeah, does that come from all of your like interactions with the Facebook folks?

[00:21:50] JM: Well, so Pete and Nick from Facebook? Well, they had worked at Facebook. Pete Hunt and Nick Schrock.

[00:21:55] CF: Yeah, yeah. Schrock, yup.

[00:21:57] JM: And they said I should do a book about Facebook. And so it took like two and a half years. And it was horrible.

[00:22:06] CF: Yes. Are you self-published?

[00:22:09] JM: We publish on our own imprint. It's published through Software Daily.

[00:22:13] CF: Gotcha. Oh, yeah. Yeah, of course. Yeah. So I think someone like you is in a very unique situation, because you do have such a large network. You do have a platform to share the book. We chose the O'Reilly folks. We've had been doing online trainings with them. I've been submitting book proposals, honestly, for eight years. Yeah, so before they –

[00:22:37] JM: What was your first?

[00:22:38] CF: Oh, man. It was probably a Spark book, because I've been poking around with Spark since 2013. But not being an expert necessarily at the time. Databricks was the main expert. So yeah, so they were getting all the book deals. **[inaudible 00:22:54]**, right? Like one of my friends lives in like San Francisco with us. She was able to get in sort of early on that whole book situation and then got hooked up with the O'Reilly folks.

Yeah. You know, man, it's a long process. I would suggest coauthor, which it sounds like you do have a coauthor, right? Yeah, those other two guys?

[00:23:14] JM: Not really. I mean, at the beginning we were thinking, “Oh, yeah. You guys are my coauthors.” In the meantime, like Pete is getting his company acquired by Twitter, and he's going to be the head of health at Twitter. Nick is getting Dagster online, which became Elemental. So like Nick is running a series A company, Pete is basically enforcing health in Twitter. And I'm like, “Hey, guys, do you have 15 minutes today to talk about the book?” They’re Like, “No, we actually don't.” So yeah, never write a book. Never write a book. And never read a book by yourself.

[00:23:56] CF: Sure. Yeah, coauthor is key. It really boils down to why are you writing the book. And some folks certainly don't do it for the money. There's zero money in this. So the reason Antje and I – So she's based out of Germany. And so we got together. We both joined AWS around the same time. We both knew each other from prior to AWS. So there was some friendship there. So we weren't going in blind. If you try to do it alone, as I'm sure you found it, you can oftentimes get discouraged. You wonder why are you doing this. What's the value? Especially talking about SageMaker and the sort of AI/ML story at like AWS, things were changing. I mean, even on as a week by week. Fortunately, I was able to hop on to the status, like email lists. And so I knew what the PMs were working on. This is not something that like an outsider, outside of the like AWS firewall, has access to. So it was very, very difficult to stay on top of things.

And on top of it, we were also doing our daily jobs. We weren't traveling as much during 2020. So that was nice. We were always at our laptops. We always had access to the cloud. And we were able to really explore how to build an end-to-end. And that was what we found sort of missing from public knowledge. And part of it was because the end-to-end story just simply wasn't there. And then part of it was it's not often clear where one product begins, or one feature begins, and then one feature ends and how to pull them all together.

But yeah, I would say, I'm a morning writing kind of person. So I would write while I was fresh and had seven cups of coffee in me. Yeah, before Polk Street would get too crowded with tons of people watching, and cars, and trucks delivering things. So that was my key.

[00:26:01] JM: So you like Polk Street – What neighborhood are we talking?

[00:26:06] CF: Bob's Donuts. So I'm in Polk Gulch, it's called. Yeah, so basically part of –

[00:26:10] JM: That neighborhood has a lot of character. I used to live in Japantown.

[00:26:13] CF: Yeah, I thought you were still in Japantown. Where are you?

[00:26:15] JM: I'm not in Japantown anymore. Now I'm in a similar – Well, actually, I don't want to say on air. But it's an area with just as much character. But yeah, I mean, I loved being in Japantown. You're adjacent to so many characteristic elements of San Francisco. You got Fillmore over there. You got like the music scene? You have a lot of his history there. But yeah, a lot of creative energy. You just feel the creative energy. If you wake up in the morning, you walk to one of these random coffee shops around Polk Street, as you said, you just feel the creative energy. Do you write on your phone? Do you write your smartphone? Do you write the book on your smartphone at all? Or you bring your laptop?

[00:26:57] CF: Yeah. I always bring my laptop. I mean, I bring it to, yeah, the Mexican place. Yeah, I bring it to the coffee shop, which is very unique to San Francisco. That's not a normal thing to do in like the whole rest of the world. But yeah, being on a laptop at a coffee shop is like totally normal. And the cool thing is, depending on what stickers you have on your laptop, I have a TensorFlow sticker, I have an AWS sticker, people talk to you, "Oh, yeah. Do you know Kubernetes?" "Oh, yeah, yeah, I do know Kubernetes." And then they just randomly ask you a question. And I can help them if possible. So people collaborate like that.

The other tip – So besides coauthor for sure, is always think about that target audience. And that was one of the hardest things was with a book this big, covering that many topics, we really needed to nail down, "Is this a data scientist? Is this an ML ops person? Is this someone that knows SQL?" I mean, there were conversations, me and like my coauthor, and the O'Reilly folks were like, "Wait a minute. Now you're introducing some SQL statements." And so there're a lot of like prerequisites that we kept either tacking on or shaving off. Yeah, that's one of the hardest things. And then use Google Docs to or some sort of collaborative –

[00:28:15] JM: Google Docs is so good. I mean, look, I'm all for Notion. But Google Docs is great. Did you have trouble getting the book down to a reasonable size?

[00:28:26] CF: Yeah. Oh, yeah. And in fact, there was a miscalculation. And yeah, this is kind of a funny story, where our editor had not ever used Google docs for this kind of thing. And so he had kind of a back of the envelope like calculation. So we would provide him, I don't know, 700 pages of Google Docs. Each chapter was its own Google Doc. So it wasn't one gigantic thing. And he would say you are only at 300 pages, when really we were at 600 pages. So he was a complete – Yeah. And so we kept writing more and more thinking that, because it's totally different on print, or in Adobe, in like PDF, yeah, final form. And then the O'Reilly folks have a very, very rigorous review process. And they were insistent that we didn't just use AWS people. And so I really liked that too. So like we brought in Sean Owen from Databricks, formerly Cloudera. Head of data science. I just pulled on all of our network. Ted Dunning from MapR, now Hewlett Packard. So these folks that had been with me, and I've helped to review their books, and they helped to review, yes, our book. And yeah, that's an invaluable part of getting hooked up with a publisher, is they have a lot of previous authors and a lot of people I really respected.

And then I would say one more tip is to keep reading. So while you're writing, it's frustrating to actually pick up a book and start to read somebody else's stuff. But, really, that kept me moving forward because. I mean, right up to the very final draft, I was adding stuff that was based or was sort of stimulated by something I had read in someone else's book. And, yeah, grabbed a sample. And then like changed it to be our like use case where we were building this like text classifier with Bert. Really, just keep reading right up to the time when you actually publish. But it's definitely not for the faint of heart. And I got a few extra gray hairs. I'm not sure if you can see them here. I shaved some of them off for you so that my like microphone wouldn't touch my beard.

[00:30:37] JM: Yeah. I mean, at this point, I've built failed companies. I've built failed projects. I built failed – All kinds of failed things. Writing a book is probably the hardest. It's been the hardest for me, especially like – So, for me, my editing process, I finished the first draft. I have an editor. I use this company called Scribe Media, which I heavily endorsed to anybody that wants to write a book. Scribe Media is awesome. But they gave me an editor, who was really

good. His name was Hal Clifford. And I wrote the first version of the book. I'm like, “Oh, God. I'm done. Okay. It’s great.” And he just gives it back to me, and it's just red everywhere.

[00:31:13] CF: It's tore up. Yeah.

[00:31:14] JM: And it was basically like you need to rewrite this entire thing. I was like, “Oh, god.”

[00:31:19] CF: Yeah. Oh, and then one other thing. We actually started with images and figures. So we were in Google Slides. We started off using PowerPoints, and couldn't figure out the whole collaboration situation there. So we ended up just switching over to Google Slides. But kind of story booking it or storyboarding it, yeah, really helped us go from chapter to chapter. And by the end of it, we had written and rewritten it, yeah, probably two or three times. But those figures really helped us out. And we ended up with almost 500 figures, which is a 500-page book on its own. So we had to go through and actually cut figures and put more into words. But yeah, is your book selling? Is it doing okay?

[00:32:05] JM: Yeah. We’re number one new release in Computer Engineering right now. It's not even about computer engineering.

[00:32:10] CF: That's hilarious. I know. Yeah, we were number one in like cloud computing, which I was

[00:32:16] JM: That’s awesome. Hey, man, well deserved. I mean, I look at *Data Science on AWS*. That seems like a very widely needed topic of advice.

[00:32:27] CF: Yeah. And with O'Reilly, you have to figure out their proposal process. And, yeah, so living in San Francisco, I have the advantage of just kind of popping into these conferences. And I actually ran into one of the like editors, the acquisition editors, Jessica Habermann. And she was doing a talk on how to write like the effective proposal to get your book in. And so I sat there, took notes. And realized when you're proposing a book, it's very similar to doing a pitch deck. And because I just gone through all that process with my startup, I knew that I had to show value prop, I had to talk about total addressable market, I had to talk

about the team, I had to say who I was, why I was an expert in this area. Yeah, talk about or like my coauthor. And you really have to break down the market. And so something that I'm picking up on here is, yeah, so you stated that this really is kind of the perfect time. People do want to know how to do data science. And it's a very, very broad topic. And we have lots of places to go. But like we knew that we wanted streaming in there. We knew that we had to cover security. Yes, any book on a cloud provider has to talk about security. So we have an entire chapter dedicated to that. We tried to make it not very boring. So we showed a lot of good examples and sort of how it all fits in. But yeah, yeah, building a solid proposal. So we give permission to the O'Reilly folks to actually use that proposal as a template for new authors. And so if you reach out to Jessica Habermann, or reach out to me, I can get you in touch with her. Yeah, she's super awesome.

[00:34:16] JM: How about O'Reilly? What a timeless brand.

[00:34:20] CF: Yeah, it was pretty obvious going with them. Mainly, they have, I don't know, two and a half million reach or something crazy. I read from like all publishers, packs, yeah, O'Reilly, like Manning, all of them. But that was the one we were really hoping for. And it's the one I had the strongest connections to. So, yeah.

And funny story, we had no control over the duck on the book. They just kind of proposed it to us and said, "Well, here's the duck." So, yeah, so that's the signed copy, right? Did I send you the signed copy?

[00:34:53] JM: You did. Thanks.

[00:34:54] CF: Yeah. Yeah, so that's actually traveled through Germany. And, yeah.

[00:35:00] JM: Why go through to Germany?

[00:35:01] CF: Oh, my coauthor.

[00:35:04] JM: Oh, Okay. Great. Oh, wow! Okay, thank you.

[00:35:06] CF: Yeah, so that thing, it's like Flat Stanley. It's been across the world.

[00:35:10] JM: What's your coauthor's background? Just give her a shout out.

[00:35:14] CF: Yeah. Yeah, Antje Barth. So she's the same role as me. Based in the European region. Comes from MapR. Lot of big data stuff, worked very closely with Ted Dunning. Yeah, Ted was her mentor at MapR. And yeah, we have a quote in there from Ted, I believe. And Jeff Bar gave us a quote, which was super awesome.

[00:35:36] JM: Well, she's going to have to come on the show also, because we're like most of the way through. We're not even like talking about *Data Science on AWS*. I'm going to miss O'Reilly Conferences. How about you?

[00:35:48] CF: Yeah. Yeah. Yeah.

[00:35:49] JM: What a sad victim of the pandemic.

[00:35:51] CF: Yeah.

[00:35:52] JM: How many life-changing relationships did you form at O'Reilly conferences?

[00:35:58] CF: Yeah. And the food was good.

[00:36:00] JM: The food was good. The food was good. And the food was good. Not as good as QCon.

[00:36:07] CF: Oh, yeah. QCon, yeah.

[00:36:08] JM: Maybe a little bit better than Reinvent.

[00:36:12] CF: Yeah, Reinvent, the coffee. I just want the coffee.

[00:36:16] JM: Reinvent is the AWS of food, absolutely. You get you get what you pay for.

[00:36:22] CF: Yeah, pay as you go.

[00:36:26] JM: But yeah. From a business standpoint, why did they kill off the conference business? I mean, they have to know. This is going to come back eventually. Is it just because – You think they'll bring it back? Is it like Jay-Z announcing the end of his career?

[00:36:41] CF: Yeah. And we actually talked about that before we started writing. We're like, "Oh, man, we're going to write this whole book and not be able to do an in-person book signing. Yeah, I mean, do you remember those?

[00:36:53] JM: Absolutely.

[00:36:54] CF: Yeah. During lunches and stuff, you can stand in line for like four hours.

[00:36:57] JM: Is there a virtual book signing platform yet? Can we do that?

[00:37:00] CF: I know.

[00:37:01] JM: Let's start that right now. That's a great indie hacker business. Spin-up an online book signing. 5.99, one-time payment.

[00:37:13] CF: Totally. Yeah, subscription. Yeah, like monthly subscription. And then we just bring on 10 authors or whatever. Yeah, my personal feeling, and I don't know anything more than you do, is that they have to bring it back. I mean, it was such a vital part. And I imagined that they didn't make that much money off of those. I mean, just speaking, because I'm sure they were very expensive, right?

[00:37:39] JM: Okay. So I interviewed Tim O'Reilly one time. And I was doing the whole brain dump thing. You know, like in Silicon Valley, where they bring the founders in and like have them whiteboard everything about the company while flattering them? What I was doing – I mean, I love Tim O'Reilly, and I love what he's built. And it's like totally aspirational for me. And so I was like, "Okay, so you basically started with conferences, and then you did books. Or wait, its books

first. Books, conferences, elearning." He said that books was pretty profitable. Conferences were way more profitable than that, and elearning is way more profitable than that.

[00:38:18] CF: Interesting. Yeah. And like we've done some of their online training. We have to, I guess, probably reengage with them there. Yeah, so recently, Antje actually posted their SuperStream. Yeah, have you seen any of their online SuperStream stuff?

[00:38:34] JM: No. What is that?

[00:38:36] CF: It's basically the equivalent of Strata, their O'Reilly AI Conference, but it's like totally online. And one of the benefits, of course, is you get people from all over the world. And don't have to find a hotel down in San Jose, which was even hard for me, right? Like living up in San Francisco, I would try to get a hotel down there. And they were all $700, $900. And it kind of opens it up too to a lot of more speakers than they normally could get, right? People that can't take off three days or two days of work, but they could just pop in for a half hour, give a talk and then answer questions and leave. But yeah, they do them. I think they have like a data one. They have, yeah, an AI one. But I really wish that they would bring it back.

Fortunately, they were able to repurpose quite a lot of those people. So they didn't have to lay off too many people, from my understanding, because I still see a lot of the same faces, but more like the online side or the marketing side. Yeah, great group of people.

[00:39:37] JM: So the online conference experience, have you ever done one of these online conferences? Do you get a lot out of it? Do you get a lot out of it relative to the in-person conferences?

[00:39:46] CF: They're a bit awkward, right? I've seen some of these platforms. I don't know the specific platform names where they have this –

[00:39:55] JM: Hop-in. Hop-in is the big one.

[00:39:56] CF: Yeah, yeah, Hop-in. I don't think that's the one that – Yeah, there was one where they had this kind of virtual lunch room thing and you would drag yourself over, and like I

couldn't even figure out how to do it. I just wanted to like talk to people even in a Zoom room or something like that. But yeah, and I do find myself skipping talks that I would normally, if I was dedicated in that space for those two days, like I would definitely go to these talks. But if there's like the whole switching theory, if I can switch off and answer 10 emails during that same half hour. But yeah, I miss the in-person.

[00:40:34] JM: Okay, anyway, if you think about data science on AWS from, let's say, the platonic ideal. Let's say you're building a company. Actually, let's say you've got Netflix, okay? Let's say you've built Netflix, and you've built it without data engineering in mind. So you've built basically awesome video streaming service that everybody loves. You've got the flywheel spinning. You got cash coming in. And you are the first person on the data engineering team. You get to basically take all the production data, all the telemetry. Let's imagine you have all the exhaust data that you could ever want. And you're building the best data pipeline and machine learning and data science workflow set possible. What does that look like?

[00:41:22] CF: Yeah, for sure. So there's this really cool segments of data engineering or the data engineering pipeline called Feature Store, right? So that's something that – And I don't know if you've had any segments so far. Have you had any segments on Feature Store?

[00:41:39] JM: Is that an AWS product? Or do you just mean, abstractly, the idea of a feature store?

[00:41:44] CF: It's both. Yeah. So there is a concept called Feature Store. And then, yeah, SageMaker has SageMaker Feature Store.

[00:41:51] JM: So this is like what Michelangelo does, right?

[00:41:54] CF: Exactly.

[]00:41:55 JM: Okay. Yeah, we did Michelangelo.

[00:41:57] CF: Yeah. And Mike Del Balso, who lives in San Francisco, started a company called Tecton, that really, really focuses on that. And I speak with him, yeah, maybe once every few

months or so. He's really, really scaling that company out. It's backed by Andreessen, backed by, I think, Sequoia also. So yeah, those guys are kicking butt. It's called tecton.ai. And so SageMaker also has something called Feature Store.

So the reason I bring that up is once you – So there's raw data, right? Then there's transformed features. And what you're trying to do is build up these embeddings. It's the classic term for it. Yes, anything can really be an embedding. You're just going from one space, from one vector space to another vector space, essentially, taking raw data, raw text. Let's say product reviews, right? Customer reviews, and then converting those into Bert vectors. And what you want to do is like just do that once. And so when you ask about the most perfect data pipeline, you want things like traceability. There are reasons now that you might need to actually take data out. So there's GDPR and all those familiar things where we might actually have to remove data even from our trained model. So at some point, it's not good enough to just delete the rows of a user from your database, right? Like you also have to delete or modify any and all models that have been trained with that user's data. This is a very, very difficult problem. It's not very easily solved. But the ability to trace and say this set of 15, or 500 models used that single person's data. We have to now pull it out and then retrain and start from that particular point in time, right? And then retrain.

And so feature stores gives you that ability, right? Where if you break it down, feature store, it's database tables, but has this notion of time that these features were actually ingested. And you can share these features just like you can share to any database. So before SageMaker Feature Store, people built these on their own. And they were sort of bespoke. But yeah, the folks over at Tecton and then the SageMaker Feature Store really tackle this. It's a really, really good topic. You should, yeah, pull someone in from Tecton or somewhere to discuss this.

[00:44:34] JM: That's really interesting to hear, that, basically, the entry point into what you would consider a sensible data engineering data science stack is this feature store. Why is that? Of all the things in the PANCAKE STACK, why is the feature store the best pancake?

[00:44:54] CF: Yeah, the best layer? Well, you can reuse, right? So these oftentimes are very complex transformations. Going from raw text to this vocabulary that Bert has learned, right? Bert has been trained on millions and millions of documents. And Bert has a notion of like

human language. And trying to get natural language text into these Bert vectors could take many days. Yeah, depending on like how much compute you have, how much data you have, all that stuff. So, you really only want to do that once or very few times, and just sort of incrementally –

And so from that point, like you could also then share not only with your teammates, but yeah – So you could use those same features that have been transformed, those same Bert vectors, when you're actually making predictions. Otherwise, you're making this like transformation twice. So you would have to do at once during training, which could take a few days. Huge, huge batch job to then create those vectors. And then when you go to actually make predictions, like you would then have to re transform that same text into that same Bert vector space. And so, yeah, really, there's a lot of sort of operational, like cost savings going on there. That's Tecton's main value prop, is they have lots and lots of cool features, but they also save you quite a lot of time. And, yeah, have a really good like value prop.

[00:46:30] JM: Okay. Well, unfortunately, we got to get going. But I'd love to keep talking to you. We should do another show. We should do one in-person maybe in a few months. I think we should just like do like a long, like the history of data engineering. What is the six-year history of data engineering? And why is it the way that it is today? But I'd also love to do a show that actually covers AWS in a little more detail if Antje wants to come on. But sorry to cut it short, man. I'd love to talk longer.

[00:46:58] CF: Yeah, man, no problem. Yes, always good to see you, Jeff.

[00:47:01] JM: You too, in-person soon.

[00:47:02] CF: For sure.

[END]