EPISODE 1409

[INTRODUCTION]

[00:00:00] KP: Protecting your customers begins with best practices for securely capturing, storing and protecting the data you collect for or about them. When an organization has a large enough dataset, needs typically arise for doing analytical workloads or training machine learning models on this data. If you use random or mock data to generate a report or train a model, you arrive at an output that doesn't reflect the true use case of the organization. Success on tasks like this seems to require production data. Alternatively, perhaps production-like data is good enough.

In this episode, I interview Alex Watson, cofounder and Chief Product Officer at Gretel. We discuss their solution for generating privacy preserving synthetic data that remains representative of the underlying dataset.

[INTERVIEW]

[00:00:53] KP: Alex, welcome to Software Engineering Daily.

[00:00:56] AW: Thanks, Kyle. Excited to be here.

[00:00:58] KP: To kick things off, can you tell me a little bit about your introduction to software and technology?

[00:01:04] AW: Yeah, my intro to software that really kind of just set me down the path I'm on today started – I think I was really lucky. I was in Indiana. Kind of grew up and started going to school in Indiana, and I had an exceptional computer science teacher that I got introduced to in sixth grade. And he started us off programming on the mainframe computer and dialing in and doing our homework assignments from our modems at home, taught us how to hack into each other's profiles and things like that, and really sent me down that path.

[00:01:33] KP: And what's the path that led you to Gretel? Or is it pronounced Gretel? Sorry?

[00:01:38] AW: It's a long one. Gretel, yes. It's a long one. And as I think about it, it makes sense. But there's a few jumps. I went to Indiana University for computer science. Had an awesome experience there. Graduated in 2001 right after September 11th. And never would have guessed this, but ended up going to the National Security Agency where I spent the next seven years, and had a just incredible experience. I got to work at the time on really kind of big data and machine learning projects that were, in my opinion, a little bit ahead of their time.

I left the NSA to go start up my own company. And what I wanted to do is actually help companies that were starting to use SAS platforms, whether it's like Office 365, or Google Suite, like identify and protect where their important data was so that they could start using those services more across their business and enable collaboration and things like that.

So I founded a company called Harvest AI in 2014. Had some incredible customers as we were building it up. We built through Series A, and then we got approached by a couple companies, AWS is one of them, around acquisition when we were going out for Aeries A raise. And such an incredible opportunity and experience there to take Macy, which was our product, and bring it to the whole world and thinking about the scale that we'd be doing in the impact. Launching AWS's first security services was really, really cool.

So we ended up getting acquired by AWS in 2016. For the next three and a half years, I was the General Manager there. Launched the first security services we had, and so much of Gretel, and our inspiration, or at least my personal inspiration for founding Gretel was around things that I saw there. A, I saw the just incredible value of access to data that you have building inside an AWS, or a Google, or a Netflix. Companies that are incredible at working with data give such an advantage. And a lot of those advantages are powered by a 500-person compliance team that many other companies can't replicate. And that creates a real business advantage. So that was kind of one piece that I think kind of leveling the playing field for any business that wants to compete in today's data world without having to collect all of the data. That's one that really hits home. And a second piece, I think, was just realizing even our most sophisticated customers, AWS. Like how hard of a time they had enabling access to data that could drive innovation. We really wanted to help tackle that.

[00:04:02] KP: Well, working at AWS means you saw some of the biggest enterprise scale problems. I'm sure you saw some unique challenges that the NSA as well. What's unique about doing business at this large scale that is especially challenging?

[00:04:16] AW: What I found during my time at AWS was I gained an incredible appreciation for scale for operational excellence. It was just something that they'd really drive really at the company culture level. There is a weekly meeting that any developer in the company is invited to attend. That goes all the way up to the SVP level. And we review the dashboards, essentially the operational effectiveness of our service. And there's a kind of infamous process where there's a wheel, kind of Wheel of Fortune style, and you spin the wheel and it lands on somebody's service. And whoever service it lands on, presents their dashboard to the entire company. And if there's a blip on the dashboard where you had an outage or you had an unexpected increase in data, you really need to be able to explain that.

And seeing that it's important all the way from a developer all the way up to like the SVP level I think was so cool for me to see that. And I think it creates that type of culture that builds really high-quality products. And that's something that we've worked to replicate at Gretel, and I think is a really interesting thing for anyone software space to look at.

[00:05:16] KP: And can we get into the core value prop at Gretel? What do you help organizations do?

[00:05:21] AW: So at Gretel, we believe that data is more valuable when it's shared. And the unfortunate circumstances, and for very valid reasons, it can be very, very, very difficult to enable access to data. Often, that data that is most valuable to a company, it's based on user information, or private data, or has lots of PII, and things like that. These things are so difficult to solve using kind of manual redaction techniques and things like that, that many companies just put up barriers and they put up friction. And while they would love to enable a data scientist or a developer that has an idea to test it out on real data, they just can't do it, because it's too risky.

At Gretel, really, our goals are we build a simple set of API's, but I think very powerful API's that help developers build privacy, or data scientists build privacy in their data. So this is a space that is gaining a lot of popularity over the past couple years. It’s called synthetic data. But to boil

it down, essentially, we take machine learning models. We train them on your sensitive real-world data.

I give an example, for a healthcare dataset, where the goal would be how do we enable hospitals to share COVID research with other hospitals. You want other hospitals to be able to learn about COVID. You don't want the hospitals to be able to learn about the patients. And that's really our goal with synthetic data, is we use machine learning to create an artificial dataset. So that's the same insights and same statistics and distribution as the real data. But it's not based on any individual person.

[00:06:51] KP: So do I lose something in that process? My model have an artificial ceiling put in place because it's not the “true data”?

[00:06:59] AW: That's an area we're spending a lot of time researching on. And the answer to it is complicated. I'll share some links for some research we just published where we took the top 10 datasets on Kaggle, the data science platform Kaggle, which are used for machine learning tasks. And we created synthetic versions of those datasets and ran the same machine learning tasks on them. No configuration, no tuning, nothing. Just kind of straight across the board.

What we saw was on average of 2% to 5% loss in accuracy that you would have. But what you gain here is a tremendous amount of privacy. So there's an advantage there. I think another area that's really exciting is there's potential for this synthetic data to even outperform real-world data on many different datasets and use cases. And that sounds crazy, right? Like how could something that's artificial actually do better than the real-world dataset?

But the neat thing about using this application of machine learning to create synthetic data is that you can correct biases that exist in the original data. You can better like essentially architect your dataset to be more fair or more representative of the type of output you want to get. So there's a couple examples. I'd be happy to dive into around the healthcare space where we've actually seen 6% to 10%, improvements in accuracy, and also fairness across different types of data when you're telling your synthetic data model what type of data to generate.

[00:08:18] KP: That's really interesting. Yeah, what are some of the measures of fairness? How do you identify and then debias a dataset?

[00:08:25] AW: That is a very tricky proposition, right? Like really understanding the second and third orders of bias in a dataset can be very difficult. There are some examples. And these actually, this foray that we had into starting with fair AI and bias and things like that actually came out of customer conversations and people using our open source synthetic data library. An example I can give you was, very early in the company, we started working with the University of California Irvine. Many medical institutions and the UC system, for example, take advantage of their hospitals to build medical datasets that they make available across, for example, the UC system.

UCI has published some like the core kind of reference datasets on Kaggle. For example, I think one that comes to mind is a disease, heart disease detection dataset. And what the UCI folks noticed was that, while this is being used all over the world to create models that detect heart disease from a set of traits or characteristics, there's biases that exist in dataset. And one that they were really sensitive to was that there was over a two to one ratio of males to females in the dataset.

So what they wanted to do was take our synthetic bottle, use it, train it on the existing heart disease dataset, and say, “I'm going to balance out. Essentially, create enough synthetic or artificial female records.” There's no substitute for having the real data, but essentially addressing the representation bias that existed in this dataset. So you have a 50/50 ratio between male records and female records.

And what this ended up accounting for was most algorithms in your training classifier to detect heart disease, they're going to get really great at detecting male heart disease. And by balancing this out, it kind of forced the algorithms to think more across both classes. And it ended up – I think if I remember correctly, it increased the overall performance of the dataset for female heart disease detection by like over 6%. And overall accuracy, the dataset across both male and female detection, by 2%.

[00:10:21] KP: Wow! Big win then. So I guess the key insight, yeah, is that you have two to one male-female in your data set. Did that have to come from a doctor who understood that, I don't know, if males are more susceptible? Or if they submit for exam more? I don't know the core cause here. But how do you draw out that insight initially before you correct for it?

[00:10:42] AW: It comes down to being a data science challenge where you have to understand the different dimensions and the classes in your data. So the UCI folks and people that are working on this dataset, I'm sure this is something they saw during that like initial data discovery phase. Other use cases we've seen around balancing income prediction datasets and things like that, even down to voter data. You see this kind of common pattern, where at some point it just becomes prohibitively expensive for you, or impossible for you to get more sample datasets for heart disease that match the same distributions, the original data.

So there's a pain point here where it's either so expensive for you to get more data to support building a better dataset, or it's just impossible, because the means used to collect the data no longer exist, for example.

[00:11:28] KP: So I'm thinking of a deployment of this. Maybe I'm at a company. We've got a bunch of good real-world data in our production system, but it's locked up. I don't have access as a data scientist. But I know it's a Postgres database, or maybe there's some Parquet file sitting somewhere. How do I plug that into Gretel to get what I need?

[00:11:48] AW: Gretel follows like a data-in, data-out source sync model. So we really don't want to be opinionated about where you store your data. For example, whether you're putting in Postgres or you're putting in as S3. From an architecture standpoint, Grete is just API's, and you submit a dataset, you submit a configuration, which is defined in a YAML-based configuration, or you can use our SDK, where we have a Python-based SDK. But you send data in. You get a configuration. You say, “I want to create a synthetic model from this dataset. I want to transform this dataset.” Really, you give it a task and say generate 10,000 records based on this training data and give it back to me.

For how you would integrate with the Postgres or an S3, I personally – And I think our team definitely recommends like the really kind of durable storage. Like, I think, the Blob storage,

whether you're using s3, or you're using Azure blob storage, or GCP storage, is such a neat design pattern, because it's so durable. You can access it from hundreds of spots at once and you're never going to impact your production database. So we see often customers, for example, taking the Postgres database, pushing updates, and essentially storing the incremental updates that they have to that database, to – I'm going to just use S3 as an example in Parquet format. As soon as that raw data hits S3, it sends a notification where you have some code, picks up an SNS notification, for example, notifies Gretel to create a synthetic version of that dataset, and then dump it into another S3 bucket. So that's a pattern we see pretty often.

[00:13:18] KP: That makes sense, yeah. And then what are my – As a data scientist who's maybe very picky about how this process is going to take place, what are my options for configuration?

[00:13:27] AW: With data sciences, we've got three different interfaces you can use to interact with Gretel. The first is the console. So Gretel's free. You can sign in from our web user interface. You can drag your own dataset in. It's really that simple. And specify like what hyper parameters you want to change, if any, and train your model right there in the UI. With data scientists, the most popular application we see between the console and the SDK and the command line interface would be the SDK.

So a lot of data scientists working inside Jupyter Notebook environments, Databricks environments, things like that, that like to use the Python SDK, because they can explore their dataset, they can make any changes to it, and really have fine tune granularity over the API and the parameters when they call.

[00:14:12] KP: And if I have a simple data set, or I know my schema pretty well, I might already be able to tell you all the columns to worry about and stuff like that. But can you help me there if I've got something that's a bit overwhelming and I maybe don't know where all the PII could lie?

[00:14:25] AW: Yeah, it’s a scary problem. And it's super time consuming for that poor data scientist or developer that's tasked with, “Hey, create a safer redacted version of this dataset.” We have the first couple stages with Gretel involved. It runs data classification. So it runs an

entity recognition across the entire dataset and tries to find where PII sits. You can run a set of default policies, for example, redact PII. Or I want you to shift these dates so they're no longer people's birthdays, or things like that, across that data set in a completely automated manner. Just give it a configuration file to find your regular expressions if you want to. But it makes it really easy to do that first completely automated paths across the data where, for example, our ML will likely find that social security number that's buried 100,000 rows deep in your dataset that you might not see just looking at it manually.

[00:15:20] KP: And I'm curious about adoption. I can definitely see a use case for like a hospital that knows they need this. And they have to push in this direction. But also, for a very forward-thinking ML ops engineer who wants to bring this in house, what are some of the typical paths to Gretel?

[00:15:37] AW: When we started building Gretel, we built it a little differently than a lot of other companies in our space. And our idea was very similar to AWS, is to give like really nice documentation, make it very accessible for you to get started with, and let this be like a developer or data scientist-driven go to market motion. And what that means for us, we have a completely open source synthetic data core that you can start working with. It's completely free to use. Our SaaS service, which is also free up to a certain level, so we get free credits every month, has a bunch of new features on top of it that make it usable at scale.

This motion, for us, kind of like learn in public and open source data in our research as we go, has really helped with adoption. So even as a two-year-old company right now, we have about 2000 developers using our platform. And we've got in the 10s of getting close to 100 companies here that are using Gretel actively, which is pretty exciting.

[00:16:32] KP: And can you cover at all what sort of industries have shown the most interest in the product?

[00:16:37] AW: Sure. Taking a step back on the industry with synthetic data, I think there's so much interest in synthetic data, and we see it like centered around several different industries. With Gretel, what we've tried to build as a platform that will work for any data scientist, any developer in any industry that wants to create synthetic data. And that's really our vision is to be

that like deep hub or the place that people go to for synthetic data. Across our customer base, we see a ton of interest. I think a lot of it driven by the need to innovate after the pandemic and after COVID in medical space. So whether it's life sciences, working with medical datasets, hospitals trying to find ways to monetize their data, all the way up to genomics research. I see a ton of interest there.

Financial space, I think, obviously has a lot of interest, and they're pretty forward leaning on adopting new technologies. But we see really interesting use case even in like, for example, the gaming industry, which has been really fun to work with.

[00:17:36] KP: Oh, what does gaming want to do?

[00:17:39] AW: I think, across the different gaming companies that we've been working with, we see a pattern where I think they've realized how incredibly sensitive the data is that they gather as a player plays a game. How the player reacts? What type of games that a player likes to play? When you look at like the more immersive VR and like AR type experiences, you get incredibly specific location data. You get, in some cases, player heart rate data, things like that. And the gaming companies, to their credit, have really looked at this as something that needs to be protected. So popular use case I could talk about right now, it'd be how do you detect bots that are plaguing, for example, many like online ecosystems. You need to train your machine learning models on really granular player data. And some of those are bots. And some of those are humans. So how do you create this to be safe, where this data set doesn't contain any single players' information? It contains statistics that are representative of the entire set.

[00:18:37] KP: And I'm curious if you can share any details on what this looks like operationally for the most mature of your customers? Is it part of everyday production, or when I need to train a model, I go and use this? What's the typical use cases?

[00:18:51] AW: We see many of our customers. That journey starts with coming in through our console or starting with some of the like open source examples that we have, for example, how to balance dataset, how to de identify or apply differential privacy to a dataset. From there, the question is how do you operationalize it? And that's where we see a lot of customers leveraging the cloud and leveraging the interconnects that exist in the cloud. So whether it's that design

pattern we talked about earlier with S3 and notification, we did a workshop using Airflow, which was really cool connecting to a Postgres database to create synthetic models of that.

So we see like a real trend once you understand what you want to do with the data, and your goal, which might be create a synthetic version of my Postgres database. Create a synthetic version of my data lake. We see a ton of interest around BigQuery, and Snowflake, and Redshift and other like data warehouses, whereas new data comes in, do you want to synthesize that data. You want to create essentially a twin of that data warehouse or table that can be like essentially given wider access to or used to train ML algorithms.

[00:20:00] KP: And I know you and I appreciate the point you'd made earlier about Gretel not being too opinionated. That companies can use it in different ways. But for people that are getting into this for the first time and really need some direction, are there any best practices you typically recommend?

[00:20:14] AW: I think I would recommend using one of those design patterns that really will enable scale in the future. So for example, rather than connecting Gretel directly up to your Postgres database, having a way to export that data going from Postgres to a durable storage that you can use however you want to. You can query it. You can scale it. Things like that work very well. So we're personally, at the company, is we're working with customers really excited about that type of automation. Frameworks like Airflow I think are also – And the DAG frameworks are really exciting for like how you can build really complex pipelines and processes like that, that are arbitrarily complex that you need. So I highly recommend that.

And we're trying to make this as simple as possible. So we are releasing – The example right now we've got, it's in beta testing, an S3 connector that you can launch just from the AWS Marketplace. So a couple clicks, it provisions all of the infrastructure that you need to solve a common data classification, data transformation use case.

[00:21:13] KP: And do any of your customers get a concern about the duplication of the data? I like the idea you described earlier about there being kind of a sibling table. But that sort of also sounds like I have two extra storage now.

[00:21:25] AW: It's really important to keep track of like how often that data gets duplicated across the business. What I would say is, really, the end product that Gretel creates is a model. And you can use that model to create whatever you want, whether it's more data, arbitrary levels of data, or you could just query the model directly. I think the advantage of synthetic data is that you have data that looks just like your production data, but it's infinitely portable. So happy to kind of dive in on some of the applications of privacy. But whether you're applying our privacy filters, or differential privacy, you get some level of guarantees around the privacy of this data that enable you to share it in ways that you couldn't do before.

So while it is, to your point, duplicating data, and you're creating another set of data, that data doesn't have the same implications if it gets accidentally emailed outside of your business, or it gets lost, or it gets checked in on accident to GitHub. So while you're creating more data, that new data has a much lower risk profile than the original data you're training from.

[00:22:26] KP: And could you give a rough definition of differential privacy?

[00:22:30] AW: Yeah, I like to think of differential privacy, really, not so much as an algorithm. But I like to think of it as a standard that algorithms can meet. And what is fascinating about differential privacy is it gives actual mathematical guarantees and protections to both kind of known attacks on privacy and even unknown attacks. So how does differential privacy work?

And I've listened to your podcast before on this topic, and I think that's a great reference, the example you did with the US Census Bureau. What differential privacy does is it insert noise into a data set that make it very difficult or impossible to figure out whether one individual person, their data exist in that data set or not?

[00:23:13] KP: Can you expand a little bit about why that would be helpful in protecting my customers' privacy?

[00:23:20] AW: It sounds so small, but if you can give actual mathematical guarantees that you cannot tell whether a single individual's data, like whether it's my data, or your data, for example, exist inside of a dataset, then you can prove that no individual users data can be compromised or linked to another data set. And the potential there, it really opens the doors for

what do I need to share data with partners or make data available across my business. You know that no individual customers data, that record that represents Alex or Kyle in the dataset could ever be proved to be part of that data set or not. It's really kind of small, but very significant guarantee on privacy.

[00:24:01] KP: So having some privacy guarantees is something I don't think anyone would be against. And we've seen an endless string of breaches and compromises. Why is this sort of the exception rather than the standard at most companies?

[00:24:13] AW: Well, we started talking about differential privacy. And there are so many cool technologies and things like that around privacy. We call them privacy enhancing technologies, whether you're looking at federated learning, whether you're looking at differential privacy, like all these different ways to protect personal information. Differential privacy is unique, and that it gives you formal mathematical guarantees on the protection of data. But what happens there is it requires an understanding of the data set that you're working with. And it also often requires like a large amount of data to work well.

So where we've seen differential privacy work in the past, or publicly we've seen it, would be the US Census data, where they're trying to protect individual census information inside of there. We've seen it with Apple. They use it for their – Kind of funny story, they use it for your keyword prediction, and predicting what emoji you're going to use. But it ends up, that it's very sensitive data, but they have massive amounts of data to train from, and they can come up with a really good model that is also differentially private.

So the short answer here is that, often, without a really large data set to work from, differential privacy causes a hit to the accuracy or the utility of the data while it gives you those privacy guarantees. This is an area that we've spent a lot of time thinking about working with our customers. And the accuracy is very important. And we see it being a balance for any customer on trying to figure out like, "How do I want to balance my privacy concerns? My accuracy concerns?" Differential privacy makes perfect sense in some use cases where you need actual guarantees. In other use cases, we've developed things we're calling privacy filters, but like always on by default, mechanisms that were designed specifically to protect against attacks on privacy that offer much lower hit to accuracy. So we actually ship these as an always on feature

by default. You can disable them, but you're looking at a 1%, if that, hit on accuracy, when the privacy filters are enabled, but really good, measurable protections against different types of attacks.

[00:26:11] KP: And would you mind expanding on what an attack on privacy really is? Is this something hackers are now doing? Or what form does an attack on privacy take?

[00:26:20] AW: Anywhere from hackers to data scientists. So that comes up. That question happens a lot. There're a couple different types of attacks that we were really careful to look at as we started building at Gretel that we see being relevant with synthetic data. So I'll list them off and maybe give you a couple examples. Maybe starting at the top, membership inference attack. So what is that attack? And essentially that's saying, given that I know something about a dataset, let's say we've got an income dataset that is released by my business, I know that I make $101,000 per year. Knowing that information, can I infer whether myself or someone else I know that information for was present in this training dataset? And it's essentially helping you me kind of confirm the belief there. Two other types of attacks, and these are actually gained a fair amount of public notoriety recently, would be either like the memorization attacks, where we've seen this with the open AI's models on GPT. We've seen it with Microsoft language models. Essentially, you have language models that can write. They're trained on massive amounts of public data, and they can write paragraphs, or blogs, or tweets or things like that for people.

Well, if you suspect that, for example, maybe this was trained on some type of data that it shouldn't be trained on, like maybe has credit card information in it somewhere. You could essentially prompt the model with my credit card number is, and then start kind of filling in the credit card number and see if the model has been trained on to memorize any credit card numbers. That's pretty scary from privacy perspective. Another kind of public example I could give there pretty relevant to the software community is with GitHub copilot, where you have these language models that are helping assisting you to write code. What if that model was trained on – And this gets into compliance area. Like what if that model was trained on, for example, a type of license that didn't enable it to be used for this use case? So that's something the model shouldn't be trained on. But you could kind of figure that out by prompting the model with your source code and trying to figure out if that model has been trained on it or not.

The final type of attack, and a really scary one, is a jointability or data linkage attacks. And there's a pretty classic example where Netflix posted the Netflix grand prize challenge. And this was a couple years ago. But it made headlines where Netflix said we're going to release a highly anonymized version of 100 million movie reviews. And the goal is for the data science community to train algorithms that could outperform Netflix's own internal algorithms. And the team that had the best result got a prize of a million dollars. So both for like the public recognition of the work that you've done, and also the monetary prize had a lot of eyes on this.

And the kind of crazy thing was this dataset was so redacted. It only consisted of a movie ID, and a date, and a user ID for the reviewer, anonymized user ID, and then a number of stars that you gave it. And by itself, that wasn't identifying. What some of the researchers that were working on this challenge realized is that you could combine this dataset with IMDb movie reviews, or things like that, where simply like the precision that you had of the date, plus the user ID, plus a number of stars could allow you to unmask that user, and then figure out every movie they'd ever rated and learn a lot about the users. And that required Netflix to take down the challenge and essentially kind of halt the challenge early.

This is a case where synthetic data allows you to create another dataset of 100 million movie reviews with privacy guarantees that there's no actual user in this data that can be joined to another dataset. So it would enable competitions like this or data sharing like this to exist in a way that's not possible with regular kind of manual de-identification techniques.

[00:29:57] KP: Great example. Yeah. Well, organizations, especially technology groups have had long histories of new challenges to fight. I remember there was a time when distributed denial of service attacks. I mean, they're still around. But now, orgs have good tech teams that know how to combat that. It seems like we're at the beginning of learning how to deal with data privacy. Would you agree? And either way, what's your take on where the current state of the art is?

[00:30:22] AW: 100% agree with you on that. And that was a really interesting example you just gave. My cofounder and our CTO, John Myers, actually worked previously at Arbor. So huge denial of service prevention platform they handle at some point. It’s like up to a quarter or a third

of the packets on the Internet that they're dealing with. And his task there was like how do I use machine learning? How do I use like this knowledge across the world to prevent DDoS attacks on my customers?

And part of what led him to Gretel was saying that, as they built this product, they spent more time building like the data cleaning and the like the data and optimization framework to make this possible as they spent actually building the product. So we view synthetic data, we view privacy enhancing technologies, as a potential, like real enabler for innovation and building products faster, rather than something that creates a roadblock that you have to go through.

[00:31:19] KP: And where does Gretel stand with respect to some of these compliance options? Can you help an organization achieve – I don't know if it's a particular compliance certificate or anything like that?

[00:31:30] AW: So whether it's like GDPR, or CCPA, we've even seen customers working on the like – There's video Protection Act, kind of these emerging standards around how different types of data get handled. We view Gretel as a tool that can help you meet data residency requirements. It can help you encrypt data as necessary to meet these objectives. We haven't built Gretel as a GDPR company or a CCPA company. There's lots of nuances. And there's lots of ways that each business kind of needs to interpret what those different standards mean to them. So we want to be the tool used to achieve it. But we aren't a service, for example, that would help you get compliance for CCPA.

[00:32:09] KP: And do you think those sorts of requirements or maybe new compliance standards will emerge as it becomes clear that technology can better protect consumers? It's just not being adopted.

[00:32:19] AW: I think it's something that consumers are starting to expect and demand. And I think when you look at Apple as an example, where they've built like such a business around the idea of respecting and protecting customer information, that I think consumers, when they're looking at products they want to buy, are going to consider the privacy and the steps that different companies that they work with every day, or products that they buy to protect their data as a real differentiator.

I can give you an example. Like I have three little girls, right? And we have all sorts of devices around our house that we use for security, and video cameras and things like that. And it's really important to me to know that that data is being used in a way that I consider to be ethically safe and responsible. So that's something that I look for and I think that a lot of consumers are going to be looking for in the future.

[00:33:10] KP: Yeah, perhaps the market can sort it out for us in that regard. Could you define that – I don't know if you've coined this or not, but I learned it from the Gretel blog, what is privacy engineering?

[00:33:22] AW: Privacy engineering is the process that we take and to enable safe access to data. And what we see privacy engineering kind of really taking off at both with our customer base. We also see big efforts around privacy engineering happening at big companies like Google and Amazon. But the idea is rather than building safeguards and walls to access data like later in the whole kind of process after your production data has been created, like actually interleaving these privacy enhancing technologies into the workflows that developers are building. So as data gets brought in, as part of your pipeline, and we talked about earlier, like some of the automated processes using Airflow and other tools like that, that you can use to create a safe version as this data is being created. That's privacy engineering. When it's built early into the process for developers, like it enables much faster access, adoption and innovation to happen later.

[00:34:22] KP: Well, I would love to think that every CIO and chief data officer out there is thinking about solutions like this and looking at products like yours. Man, I hope that's true. That's probably a common path. But given the accessibility of Gretel, I think there's also the opportunity for this to come bottom up that software engineers and developers could bring these sorts of technologies in house. Can you talk a little bit about what it would take for someone to do that? What's the standard getting started path or the Hello World?

[00:34:50] AW: Yeah. For us, we've got a guided experience in the console. So all you need is a GitHub ID or a Gmail address currently, even a company email to sign in. From there, what we tell developers is try building synthetic data by itself. Sounds scary. You haven't worked with it

before. There's a big potential time investment. What we want to do is like make that as easy as possible. So with Gretel, when you sign in, you're provided a couple of sample data sets. And you can choose or drag your own.

And what I would encourage any developers out there to do is try creating your own synthetic model. Try creating your own static data and just kind of go through that process. So that when you're working with data and you find, “Hey, I wish I had a safer version of this production data to work with, or I want to train a machine learning model.” Like this becomes a tool that you can adopt really quickly to go through there.

Second thing I would highly recommend, and we see a lot of our customers using, there's such great online resources for getting started with synthetic data. Actually, we do a fair amount of publications on Towards Data Science, which is a popular data science platform on Medium talking about differential privacy, talking about how you can solve a problem you might have at your business. So whether it's our blog, Towards Data Science and things like that, we've got open source freely adoptable examples you can use that are really based on the pain points we've seen with different customers.

[00:36:11] KP: We'll definitely have some links in the show notes that listeners can follow up on. Maybe to wind up, can you tell me a little bit about recent releases or anything exciting that's coming down the pipeline for you guys?

[00:36:22] AW: Yeah, at the end of this month, we're actually launching general availability. So what does that mean? That's very kind of Amazon, AWS type term to use. And what that means is we're going from our open beta, where we were spending a lot of time learning from our customers having tons of discussions and trying to figure out like what are these workflows that you want to solve for your business? We've built those workflows. And we've been working behind the scenes to enable scale. And really, for us, GA, in addition to having the public pricing and things like that as part of it, really means that we have scaled synthetic data. We've scaled data classification, data transformation, those kind of core parts of Gretel, to be able to support even what some really large customers we're working with are doing at operational scale.

We've introduced the ability not only to run as a SaaS service, but actually to deploy our API's, our workers as containers inside your own environment. So if you're working with really sensitive data that cannot leave your VPC or cannot leave even your laptop, for example, you have the ability to take all of the capabilities of Gretel, including the synthetic model training and creation, and deploy it as a container inside your environment and task it and build there. So those are a couple of features we're excited about coming down the pike. And this is an area we've been blogging about quite a bit recently. We see a lot of interest in in customers kind of expanding that notion of synthetic data from text and tabular data into other formats, whether it'd be images or audio, video simulations, things like that. And that's an area that we're actively working on. And we're creating open source examples to start with that we're pretty excited about.

[00:38:02] KP: Very cool. I'm eager to follow that up and see where things go. Alex, thanks so much for taking the time to come on Software Engineering Daily.

[00:38:08] AW: Thanks, Kyle. Appreciate it. It's a great conversation.

[END]