SDS PODCAST

EPISODE 1030: GARBAGE IN, GOSPEL OUT: WHY AGENTS NEED BETTER DATA, WITH SALESFORCE'S GAURAV PATHAK

Jon Krohn:	00:00	When humans get bad data, they argue about it in meetings. When AI agents get bad data, they pick a number and present it with total confidence. My guest today calls it garbage in gospel out. Welcome to another episode of the Super Data Science Podcast. I'm your host, Jon Crohn. Today's guest is Gaurav Pathak, SVP of product management at Salesforce. Gaurav spent 13 years at Informatica building its metadata and AI products, including the Claire AI engine before Salesforce acquired Informatica last year. In this episode filmed live at Dreamforce in San Francisco, Gaurav explains why context is 95% of the battle for enterprise agents, who the sin eaters are that pay for an agent's mistakes, and the three skills that matter most for AI engineers today. Enjoy. Gaurav, welcome to the Super Data Science podcast.
Gaurav Pathak:	00:52	Thank you for having me, Jon.
Jon Krohn:	00:54	Love to be here. Thank you for inviting me to Dreamforce. We're filming live from Dreamforce in San Francisco, and it has been an amazing day so far. It's my first day ever at a Dreamforce.
Gaurav Pathak:	01:04	Oh, wow. How's the conference been so far?
Jon Krohn:	01:06	It's been amazing. The biggest thing for me was Gwen Stefani played before Mark Benioff, the Salesforce CEO, spoke at his keynote, and she's been such an important person to me my whole life. I've never seen her perform before, and she played Don't Speak, which is this heart-wrenching song. I literally burst into tears in public.
Gaurav Pathak:	01:25	She's perfect. Yeah.
Jon Krohn:	01:27	She is.
Gaurav Pathak:	01:28	Look forward to the full performance.

Jon Krohn:	01:29	Yeah. And there were other people, I was going to say lesser known people, but these days they're almost as famous. Sam Altman's here. Dario Almadeo is here. Obviously Mark Benoff is here. Matthew McConaughey, Reese Witherspoon.
Gaurav Pathak:	01:43	Almost as famous. Take
Jon Krohn:	01:44	Them.
Gaurav Pathak:	01:45	Yes. Great. I love the Dario's interview in the morning as well. And then Jensen was great as well.
Jon Krohn:	01:51	Jensen won. Oh my God, I forgot him. Yeah, maybe the biggest kind of hitter of them all. And it's just there's so many. How can you even remember them all? Yeah, they were really funny. Jensen and Dario were both really funny. The banter with Mark, you can tell they know each other really well. Anyway, we're not here just to talk about Dreamforce. We're here to talk about some meaty technical stuff for our listeners. So you spent 13 years at Informatica building the metadata and AI products there, and now you've been inside Salesforce for about 10 months post acquisition. Does that mean it's your first Dreamforce?
Gaurav Pathak:	02:24	It is my first Dreamforce as well. Loved all the performances, loved all the media sightings and celebrities as well, but Dreamforce is at such a different level of being a software conference. This is the OG of software conferences. So it's amazing to be here, see all the people who are participating and see all the core problems that they bring in, like how to be an agent take enterprise, how to get that data ready for AI. We love those interactions.
Jon Krohn:	02:53	Sure. And speaking of problems, what changed about the problem that you are solving when you went from being

at Informatica, a data company, to being acquired by Salesforce, which is now an AI agent company?

Gaurav Pathak: 03:05 So a lot. And it's not just the acquisition that changed the things. It's also what's happening in the market with AI coming in at full force. We at Informatica, when we started doing the metadata work all the way back three decades ago with products that were very technically focused for data engineers and such. So being able to extract metadata, which is nothing but the example that I gave in my talk is you imagine going to a supermarket and then you all see our tins without any labels on them and you have to buy a tomato soup without. And if you're expected to open each of them and smell them and then get the tomato soup, that's going to be very, very difficult. Metadata is - It's a

Jon Krohn: 03:54 Hygiene nightmare as well.

Gaurav Pathak: 03:56 Oh, if they allow it and then so on. Absolutely. What metadata really is the label on these packs and cans. It tells you what this is about, what are the ingredients, what it is good for, what it is not, and so on. We were in the business of collecting this metadata about every data asset in the enterprise and making it available for humans. We created the first data catalog way back in 2010s. We called it the enterprise data catalog. It was the Google for enterprise data assets, one place where people can come to find the most relevant trusted data asset for their analytics needs or for their data science needs. But coming into now and then a year that has gone past, we are now seeing more and more AI agents that need the same data with a lot more explicit metadata about what's in there.

04:57 Because as we have seen these agents getting deployed into the enterprise, they are seeing these data assets. It's like the Indiana Jones warehouse, I don't know, I'll date myself, Raiders of the Lost Arc big warehouse with alien

artifacts and things like that. And how do you find the alien artifact? It's like that for agents. If you don't have these labels, it'll be every can to be opened, to be smelled, and then you get to the tomato soup. With metadata, you get to find it a lot easier.

Jon Krohn: 05:26 So it sounds like in the agent world, if you had to be inspecting every grocery store can to find the tomato soup, that would be a lot of wasted tokens, I imagine.

Gaurav Pathak: 05:35 Absolutely. So that's the other thing. You'll spend a lot of time to first find where this thing is. You're looking for your customer data, you're looking for your supplier data, you're looking for a tomato soup data. And it's very, very difficult because you have to go and look around every database. We have worked with customers who have 800,000 data stores that PepsiCo is a good example. And if an agent has to do that again and again in every database, the amount of tokens that will be burned will make every AI company happy.

Jon Krohn: 06:09 And it seems like having bad data in this agentic world that we're now in is a far bigger problem than if say the data were just flowing into dashboards or something like you might have done a decade or two ago. Do you want to tell us why bad data are a bigger problem now than ever before?

Gaurav Pathak: 06:31 So the difference I call between these two things is a data brawl that used to happen earlier versus - A data what? A data brawl. Data

Jon Krohn: 06:40 Brawl.

Gaurav Pathak: 06:40 That's right.

Jon Krohn: 06:42 Oh, B-R-A-W-L,

Gaurav Pathak: 06:42 Like a fight. That's a fight. And I'll talk about how that is versus a silent failure, which is happening now. In a

decade before now when we gave every employee in an organization a self-service BI tool like Tableau or Power BI, that opened up a lot of opportunities for these employees to now work directly with data. But the problem was, again, the same thing, which data is trusted, which data is of high quality. It was very common to go in a meeting where you'd have some metric that you're looking for, number of employees in my group and three people coming with different numbers. A sales guy will have a different number, a finance guy will have a different number, the org guy will have different number, and we would call it the data brawl. So affectionately that we talk, oh no, something looks wrong with your number and so on.

07:31 And then you'd figure out what really happened. People got the wrong definition, people got the wrong data, some data quality issue that happened. The good thing with the data brawl though was that it was loud. So we knew that there was something wrong. What's happening now is the agent picks one number. It brings that number to the human. If we are in a dashboarding case and it very confidently tells the human that this is the number, we call it garbage in gospel out because now these agents are so persuasive that they can tell you anything and you will believe them. But even worse is cases where these agents are making these decisions themselves. They look at the data and they decide which customer to call based on number of calls logged, for example. And in those cases, humans are even missing out of the loop.

08:27 And if they are not getting the right high quality data, they make wrong decisions that land eventually at the doorstep of a human who would have a very hard time explaining the decisions.

Jon Krohn: 08:39 Garbage in, gospel out, that's going to stick with me for sure. So given that with what you're saying, it's crystal clear that having the plumbing right, having the data

quality right, an Informatica-like backbone is essential to being able to have an effective enterprise agent system today, especially when you can have lots of different agents from different providers working each department in a business. Or you could imagine even within one department, like a data science team, they might have multiple different agent vendors. And so you need to have this single source of truth for metadata, for data. So it sounds clear to me actually to get agentic AI right in an organization, you need to be making a big investment in the data.

Gaurav Pathak: 09:27 Absolutely. That is the thing that enterprises bring in to intelligence. It is their business. It is all of their data about the business, which is the context that feeds into these agents themselves. And it's very, very important for enterprises to get that right. You're completely right that there will be a lot of model choices. We have seen large organizations where it's almost become a wild, wild west. People download models from Hugging Face, from other places. And in some cases, these are not even approved models to be available to employees, et cetera. And governing that has become a big challenge itself. You don't want to be sending data to model providers and then vendors who may use it for training the models or even worse things as well. So governing that, making sure that your data is in the right place with all the labels properly put in is very, very important for organizations to invest in right now.

Jon Krohn: 10:24 And so every CIO, CEO, CTO, they want to be spending money on agents. Probably they feel like that's the sexiest thing that they could be spending the money on. How do you convince them? How do you make the ROI case, the return on investment case for good data management when the big shiny thing that they want to be spending on is the agents?

Gaurav Pathak:	10:46	That's a great question. I mean, that really describes my job there, Jon. The good news is -
Jon Krohn:	10:54	It's a good thing you're so shiny.
Gaurav Pathak:	10:57	The good news is as these enterprises, as they have invested in these agents and then the models, what they have realized is that just investing in the intelligence is not good enough. We have seen examples of companies where these agents have been made available to their customers like support bots and things like that you see on the websites. And this very soon realized that all the benefits that they created this agent for are actually backfiring. Example, in case of support chatbots, they tend to escalate a lot more back to the humans than that was happening before. They would take users down the wrong path, provide guarantees that the organization cannot meet. All of these come back to a human. And then we have worked for those humans who actually have to pay for the sins of the model. These are the sin eaters. So eventually model does the sin.
	11:58	Humans who have deployed the models are the sin eaters and who have to pay for it as well. So organizations in 2026 realize that how important it is to give them right data set. There are now tools available that can show when the model is going wrong. We are creating those tools in Agentforce 360 that was announced today, the enterprise AI harness and giving users that transparency. Okay, this is the case where model does not even have information to answer properly. Can you give them the right data sets? We are providing all these kinds of tools to organizations to be in the better.
Jon Krohn:	12:31	I love that. And then so basically what you're saying is, because my question was kind of how do you make the ROI case? And basically what you're saying is that in 2026, it's maybe not as hard to make as it used to be.

People are realizing that there are data gaps and they need to be spending money on getting it right.

Gaurav Pathak: 12:45 Absolutely. Just investing in agents and then the only intelligence is not going to be the case. Most organizations are realizing context is about 95% of the battle. 5% is intelligence and then choosing, but most of the battle for an enterprise is to get the context right.

Jon Krohn: 13:03 Love it. While researching for this episode, I found this stat that you seem to go to recurringly. You say that customers used to write three to four data quality rules in a good week, and now they use something called Claire to generate around 200 of these data quality rules per day. So first of all, what exactly is a data quality rule? Because I'm a data scientist, I've been doing this for a long time, but that isn't even something that's an obvious term to me. I guess it's kind of like a data management term that isn't my forte. So what's a data quality rule? How were customers writing these before and now how has it gone up two orders of magnitude in terms of the quantity that they're making?

Gaurav Pathak: 13:45 Sure. So a data quality rule, you can think of them as automated pieces of code that check whether the data that is feeding the agents or feeding an analytics model is of high quality. Basically

Jon Krohn: 13:57 Those kinds of flags you were talking about in your previous answer where an alarm goes off and says we don't even have the data for this kind of question.

Gaurav Pathak: 14:03 Exactly. So things like that. For example, let's say we are looking at profit and then we have the revenue data and we have the cost data. Profit should always be revenue minus costs. So you can give this as a rule to automated system that always checks for all the entities that we have across the world. Profit should always be revenue minus cost. There may be some cases where it is not and now

the data quality rule fails and says there's something wrong with this data value over here. Now do this for tens, sometimes hundreds of millions of these data elements like profit that are in the enterprise number of customers, numbers of employees, all these different metrics and KPIs that organizations track. And for each metric, you create 10, 20 different data quality rules because you are looking at all the different dimensions of it as well.

14:57 So now you have large number of data quality rules to manage as well. So what we've done with Informatica's Claire, which is an AI engine that we actually launched way back in 2018. At that time there was no generative AI, so it used to have machine learning algorithms and things like that, but now it works off these technologies and agents as well. You can now ask it to generate a data quality rule for profit. You can give it subject matter expertise in terms of, oh, profit should always be revenue minus cost, and it generates the data quality rule. It generates test data to validate it. It makes sure that the data quality rule is okay and validates it like an expert human would. And it can do it with an Excel file with thousand data quality rules that you uploaded in an hour or so.

15:45 So we are always making it a lot more productive for users to use Claire.

Jon Krohn: 15:50 Completely new world, which is interesting given that Claire is eight years old. It sounds like something really novel, but you were figuring out how to work the kinks out of it a long time ago.

Gaurav Pathak: 15:59 Absolutely. When we launched it, one of the biggest problems was, and it's still a problem, is I don't want to be keeping my sensitive data on these analytics platforms in public outside. I need to be scanning them and making sure that users' SSNs are not on some file that I uploaded

to my S3 folder and things like that. So the machine learning algorithms that we added in Claire all the way back then was to scan these things and say, oh, these looks like SSN, automatically mark them and then classify them so that users have this. Now the same thing, you don't want to send these SSNs and credit card numbers and God knows all the sensitive data to an agent who can use it in various different ways that we do not know about.

Jon Krohn: 16:42 Right, right. Yeah. Social security number is not the thing we want our agents to be processing for sure. And yeah, so Claire, I'll have a link to that in the show notes, but it's C-L-A-I-R-E, and I suspect the AI is part of what made that such an attractive name choice. For data scientists listening, our core listener is a hands-on practitioner today. They're very likely to be an AI engineer actually. But for those folks listening, what kind of skill or habit now matters more that we're kind of in this agentic era and that agents are consuming so much more data than people are?

Gaurav Pathak: 17:21 It's a great question. I would say three things. As an AI engineer, one should know completely about how evaling these models happen, extracting traces from what evals have been created to understand where these models are failing and making them better. The goal should be that you're creating an automated system that improves on its own. By looking at creating these first set of evals, looking at where the models are going wrong and getting it the right data, you can set the model up for that path. Second, we talked about data itself quite a lot. Data is the lifeblood of all of this, making sure that the right context reaches the right model when it has to answer a question. And third, very, very important is also the economics of it and it's not just the accuracy of it, but also the economics. What is the best path for the agent to be able to answer the questions that I'm creating the agent for?

	18:21	And then that may mean that it should not open up all the labels. We label the right things so that it reaches to the right data sets very, very fast as well. So the token economics is going to be another big thing that AI engineers need to look about. Great
Jon Krohn:	18:36	Answer. Yeah, you're about to reel them off.
Gaurav Pathak:	18:40	Perfect. So the context is king and then it's very, very important to get that right evals and traces. And number three, to be able to do token economics better. So that is - Love it. Yeah. I can't
Jon Krohn:	18:54	Believe you had those all just ready off the top of your head. Gairov was not prepared for any of the questions that I asked him today. We basically, he had just enough time to sit down and us hit the record button, which means that I didn't get to warn you for my penultimate question, which is always the same. And it's, do you have a book recommendation for us?
Gaurav Pathak:	19:12	Oh, that is a great question. A book recommendation for data scientists and engineers. It can be underneath it. Okay. I would then give you such an esoteric book there, Jon. So it's a book called The Mind Illuminated.
Jon Krohn:	19:28	The Mind Illuminated.
Gaurav Pathak:	19:28	Yes. It's like a meditation book. It's for all the times that you are overwhelmed with all the takeoff of these AI and AGI models that are happening right now. You get overwhelmed by it. Mind Illuminated will teach you how to calm down, get the light back into your mind as well. So it's just amazing.
Jon Krohn:	19:50	I need that. I need that for sure. I do have a daily meditation practice and the days that I actually spend half an hour sitting doing it as opposed to. I'll do it every day, but sometimes it's five minutes while walking the dog or something.

Gaurav Pathak:	20:04	I'm hoping these AI models, we get a lot of time walking the dogs and doing meditation as well.
Jon Krohn:	20:10	I can't wait. And then my final thing is, so how should people follow you after this episode? You gave a great interview, really enjoyed hearing from you. How can people follow your thoughts in the future?
Gaurav Pathak:	20:18	Oh, I'm on Twitter and on LinkedIn. Please search for Gaurav Patak from Informatic or Salesforce, and you'll be able to find me easily.
Jon Krohn:	20:25	I'm sure we'll have it in the show notes as well. Garov, thank you so much for taking the time out of your busy day here at Dreamforce 2026 in San Francisco. It's been a treat.
Gaurav Pathak:	20:33	Jon Krohn, same thing here as well. Amazing questions. Thank you. It's good talking to you. Thank you. Awesome.
Jon Krohn:	20:40	Great episode today. In it, Gairav Patak detailed how metadata are like the labels on supermarket cans. Without them, an AI agent has to open and smell every tin to find the metaphorical tomato soup, burning tokens across what can be hundreds of thousands of data stores in a large enterprise. He talked about why bad data used to cause a loud data brawl with three people bringing three different numbers to a meeting, whereas today an agent picks one number and delivers it with confidence. That's garbage in, gospel out. And finally, he provided his three priorities for AI engineers. One, evals and traces so that systems improve on their own. Two, getting the context right to the right model, and three, mastering token economics. I hope you enjoyed the conversation today. To be sure not to miss any of our exciting upcoming episodes, subscribe to this podcast if you haven't already.

21:31 But most importantly, I hope you'll just keep on listening. Until next time, keep on rocking it out there, and I'm looking forward to enjoying another round of the Super Data Science podcast with you very soon.