# The Distributed Data Show Podcast Episode #136:

Jesse Anderson has 1 TRILLION dollars to make your data team better - TRANSCRIPT

Release Date: February 19, 2020

Patrick McFadin: Hey everybody, Patrick McFadin at the Distributed Data Show. We're live at ... well, live. You're going to hear this recorded, but we're at Data Day Texas, one of my favorite annual events. I'm here with Jesse Anderson from the Big Data Institute. Hey Jesse, how you doing?

Jesse Anderson: Doing good. Glad I bribed you to get on the show.

Patrick McFadin: Oh yeah, you gave me, was it $10 billion? And I'm not kidding. If you had a camera, you could see this. But this is a very interesting $10 billion because where is it from?

Jesse Anderson: It's from Zimbabwe. So it's very motivating to give that $10 billion out to people.

Patrick McFadin: Well, when you said I'm going to give you $10 billion, I got kind of excited and then I realized I'm not excited anymore, but this is actually really cool. Interestingly enough, you just gave a really cool keynote about data teams, and you have a book coming out about data teams, which I think is going to be really helpful for a lot of folks out there, but maybe talk a little bit about, what is your book about?

Jesse Anderson: The book's about exactly that, data teams.

Patrick McFadin: That seems very good name then. Data teams, right? Yeah. All right, yeah.

Jesse Anderson: It's about you don't just need data scientists, and that's a common misconception. I'm going to go do something, let's say with Cassandra, with Distributed Systems. I'm going to hire a gaggle of data scientists. It doesn't work, you actually need the entire data team. You need data engineers, you need operations, because the three of them bring value all together, and when the three of them are altogether represented correctly in correct ratios, you can create this incredible value. You can take that $10 billion of defunct currency and actually make that a real $10 billion.

Patrick McFadin: You had a really good point in your talking. You said, What if I gave you $1 trillion?" I thought that was really cool, because everyone was like, "Yeah," but then you pull out that stack of Zimbabwe dollars. You were making a really clean point. What was that point you were making with that?

Jesse Anderson: That you can have a stack of $1 trillion, $2 trillion, but if it's not worth anything, if you're not creating value with your data, then it's a pile of cash. It's a pile of nothing. It's a pile of defunct currency. You can have all your data, but unless you're creating specific business value... One of the things I talked about in the talk is, there's basically four levels that I've seen of data or value creation in businesses.

Jesse Anderson: There's this extreme value, where if I went up to the business and I said, "Hey, we're going to cancel your data projects," the business would throw up their hands, start screaming at you saying, "No way, absolutely no way." But then there's another level of stagnation, or creating very minimal amount of value. That's where the business says, "Meh, whatever." I mean, we're throwing money at it. I'm just going to stop throwing money at it.

Jesse Anderson: Then there's the creating stagnate in where we're stagnating. Where if you talk to the business about it, they say, "What projects? What business? What business value?" Where when you say, "Hey, let's cancel that." They'll say, "Okay, I'm not even using it. What difference would it make?" That's where people who are listening to this show need to think about it. They really need to think about what value is it creating? Is it creating this extreme value or this stagnant or low value? Those are the projects that have a real possibility of being canceled due to lack of value creation.

Patrick McFadin: I don't think you just came to this conclusion just one day. You've been in this industry for a long time, we both have, and I remember distinctly thinking, at one point, it's like I feel like we're creating a lot of things just because they sound cool, or because Facebook did it, or Google does it. So we need to emulate that.

Patrick McFadin: First of all, there's this moment of irrational exuberance about, "Oh, we must analyze all the data," but where do you think we are in that curve now as an industry?

Jesse Anderson: I wrote a post about it, and I talked about, this is one of our dirty little secrets. We create really low amounts of value. We see that Facebook's creating massive amounts of value, and then we have our dirty little secrets of, "Hey, at my company we're doing big data on the sly. We're doing it, but we're not really doing it. We didn't really need that Cassandra cluster. I wanted that on my resume." We're in this really gray area. Yeah, we're off the hype curve, but some companies are still hung completely on the bottom of the hype curve saying, "Oh yeah, I want that distributed system. I want me that CAFCA," but you didn't need it and you've just wasted a bunch of time and effort.

Patrick McFadin: Yeah, I did this a long time ago, but do you remember the underwear gnomes from South Park?

Jesse Anderson: Question mark, question mark, question mark.

Patrick McFadin: It's like, step one, collect all the data. Step two question mark, question mark, question mark. Step three, profit. I don't know, I can understand. I mean, someone has created value of data somewhere. That has happened.

Jesse Anderson: Oh, it does happen.

Patrick McFadin: There's the obvious ones like the Facebooks, and the Googles, who've shown us, okay. Data is kind of important here, but what's a good example of a company that has used data really effectively? You don't have to name names, but what's value? What is value in data?

Jesse Anderson: What is value in data? It's taking a 'can't' to a 'can'. That's my opinion. It's a business can't, so let me expand on that. My definition of big data is can't. It's not one of the five B's, three B's, four B's. It's, I can't do something. If you're a technologist and the business person comes to you, or the business comes to you, and says, "Hey Patrick, I want to do this analytic," and you curl up into a ball and say, "I can't do that. My data warehouse will just crumble." That's a can't. Then there's an inherent business value of, there was a reason they came to that. They weren't just trying to put you to work or keep you busy. There was some business value to that. The business value that you create is done based on taking a can't to a can.

Jesse Anderson: The companies that I've worked with, that I've mentored, they aren't able to turn things around fast enough that they aren't able to actually go through and say, "What is the market actually wanting right now?" By taking that from a can't, or we have to do that every month, to every day or every week. Now the reaction there is an increase of sales of 10 to 20%. Or another one that I can talk about, because I present it in a Flink-forward, is Airbus.

Patrick McFadin: Oh yeah.

Jesse Anderson: I work at a thing with Airbus where we took a can't, that there wasn't a dataset out there of real time airplane where every airplane in the world is. That data set didn't exist in real time. There was historical ones, but there wasn't one of every single plane in real time being ready to go and be available. By taking that can't to a can, that enabled Airbus to create all these other analytics, real time analytics. This is what is going to change

the next level of, we talk about environment, we talk about issues, we still have to travel. We still have to get on the plane, but what we can do is we can make those planes more efficient, not just by making the engine more efficient. That's one thing, but you know how inefficient that go-around is and holding patterns are? They're incredibly inefficient. The plane is still burning that fuel. Let's eliminate that.

Patrick McFadin: I had this visceral reaction when you were telling me like, "Ooh, that sounds like a cool project. I want to do it." But then, I really was not thinking about the value. I'm thinking the technical problem of collecting that data, analyzing it, and storing it, is really fascinating. I think that is a good thing. A bad thing about data nerds, data engineers, is they get excited about the problem, but it sounds like you found the solution. Like okay, there is actual value in that, but beyond that just raw passion for like, Ooh, that's exciting. What's the most important feature of a data team?

Jesse Anderson: If we do it, answer it. Skills wise, it's having distributed systems and programming. Being in data engineering team. For a data engineering team, sometimes companies will take their data warehouse team and say, "Boom!" You know, take their magic wand and boom, you are now deemed a data engineering team.

Patrick McFadin: Poof. Yes, I've seen that.

Jesse Anderson: Yeah, that doesn't work. That's how companies start generating no value because the team that they dawned, that data engineering team, if they don't know how to program, if they don't know how to have that distributed systems, guess what? They're going to go nowhere, and so foundationally, if you lack those two things in your engineering team, you will go nowhere. You have a 1% chance, in my opinion, of taking that data warehouse team and saying data engineering, that's where big data projects go to die.

Jesse Anderson: Yeah, in the back room.

Patrick McFadin: In the back room. It's let me hold onto this. Let me run this as if I was a data warehouse, because you and I have talked to a lot of data warehouse people, and we've taught quite a few, and there's a very difference in mindset mentality around, Okay, I've spent 20 years trying to keep people from running a query that's too big and don't let exhaust that memory. We're like, Oh yeah, knock yourselves out. You want to do a query of all your petabyte of data? Oh yeah.

Jesse Anderson: Exactly.

Patrick McFadin: They do that because Andrew can do that. There's this whole mindset structuring change that has to happen. Otherwise, the team will always carry that on and saying, "No, no, don't do that. Don't do that." When I mentor a client, or I work with a client, that's one of the things I'll have to reiterate and say, "No, you're thinking about the problem wrong. You're thinking about this 'scarcity of resources'. You don't have that scarcity of resources with these distributed systems anymore. It's a function of costs, definitely, but it's not a function of a can't." We've now transformed that can't to a can.

Jesse Anderson: One of the things I thought was really interesting in your talk this morning, which I think really needs to get absorbed by a lot of people, is that a data team cannot be in the back room. They have to be much closer to the top, because it's so aligned with the important company values and I think that's a missing element. I mean, would you say that that's a clear indication of success or failure, is how close they are? I think you even gave an example, like you should be in board meetings.

Patrick McFadin: Yeah, yeah. To what you were just talking about, does your CDO, do you have a CDO? Do you have somebody at a C level that is representing the team? That's a really important part that I'm seeing as a clear way that companies succeed, that there is a representation. This isn't the data engineer backwater. I saw something about data scientists in CIO Magazine, we need that. Rather, the business has continued. They didn't stop at the data strategy. It's both sides having this problem. Sometimes the business will say, "We've created the data strategy, washed my hands, go knock yourselves out nerds." Then on the other side you'll have your nerds, your data engineers, saying, "I don't want to interact with the business. The business is kind of weird and I have to talk to people."

Jesse Anderson: They're suits.

Patrick McFadin: They're suits.

Patrick McFadin: I've heard that so many times. They're suits. Yeah, but you've got to get along with everybody.

Jesse Anderson: Well, here's the difference between software engineering and data engineering. Is that yeah, in software engineering, you should have been working with the suits, but you could get by with it. With data engineering, if you're not working with the suits, you're not creating something that businesses can actually use.

Patrick McFadin: That's what resonated with me. I was like, yeah, if you're really creating value, you need to be in that conversation. You can't just create technology to create technology. We're just burning infrastructure at that point.

Jesse Anderson: If you're a visual person like me, remember the wall, the ceiling of the Sistine Chapel? There's God touching Adam, and that's the, how close to the business problem are you? Are you touching each other? Or are you so far apart that you didn't even recognize that person there's such a chasm between you? In order to create the highest level of useful data product, you're touching.

Patrick McFadin: Yeah, you can't be a weekly report that no one reads.

Jesse Anderson: Yeah, just be careful because, human resources may not like that touch.

Patrick McFadin: That's a physical business metaphorical metaphor.

Jesse Anderson: Metaphorically touching.

Patrick McFadin: Just keep it, even metaphorical...

Jesse Anderson: Keep it platonic people.

Patrick McFadin: That's right. Nothing is actually being touched in the making of this podcast, just got to make sure we know that.

Jesse Anderson: Although Patrick was touching me, and I felt uncomfortable.

Patrick McFadin: But you gave me $10 billion, I felt like we had a moment. Switching gears a little bit away from data teams a bit more, you and I had some really fun conversations about technology. I think between the two of us, we know everybody, and you're getting nervous all of a sudden. I can tell. You're like, don't, we're not going to go there. This isn't a shade talk. I want to talk about the more positive stuff. I think there's really cool stuff happening, especially in the world of data engineering and data analysis. Anything, data. It always is a moving target. What's some particularly interesting things that are happening now that you think are not being noticed?

Jesse Anderson: It's starting to get noticed. I realized the need for this realtime thing that we're doing with CAFCA and Pollstar, and I started doing that five years ago. Five plus years ago. And I realized this because I was out in the field talking to people, seeing this need. CAFCA has become pretty popular, but this whole notion of what is realtime actually mean? What does real time

do? Sometimes we just focus on, Oh, I can create a real time dashboard, I can do this ETL in real time. I think the actual bigger difference is that, now when you go to pipe things around in the right places, to the right database, Cassandra, for example, before we'd have to say, "Okay. Oh God, this is going to be a terrible thing. We're going to have to do this, and we're going to have to do that, and we're going to have to do this." Now we just say, "All right, we're going to consume from this topic and we're going to pop that in Cassandra and we're all good."

Jesse Anderson: You wash your hands of that, and you say, "Okay, next database. Okay we're going to put it in here." The sorts of changes that that brings about is a ability for the data engineering team to quickly react, to be much more agile, and for the data scientist, as a direct result, to be even more agile because they get the right tools in their hands faster. But we got to rewind, you got to reset that back and say, "I disagree with some of the notions of everything should be real time." Well guess what? There are things that benefit from real time and there are things that don't, and you should know things like your SLA's are going to increase. Your difficulty on your data engineers, primarily and to a lesser extent your data scientist, is going to increase. It's going to put some stress on your organization where, if you were barely getting by with batch, guess what? Real time is even more difficult and it's really...

Patrick McFadin: That's a really high level of knowledge, and that ups the game for everybody, from infrastructure to even software engineers, is dealing with realtime data requires a whole different quality set. Like I say, when it goes into the data Lake, you can be a data procrastinator. We'll figure it out. When it comes in real time, you got to deal with this now, and you got to deal with the flow and, the thundering herd. Like, "Oh, Black Friday events. Can you handle that?" I think that's the gold standard of can you handle it? Is Black Friday.

Jesse Anderson: Yeah, and for retail...

Patrick McFadin: Cyber Monday.

Jesse Anderson: That's actually an interesting question that I do. When I work with a team, that's one of the first things I ask them. What's your high watermark of the year?

Patrick McFadin: What's your Black Friday?

Jesse Anderson: Yeah. What's your equivalent of Black Friday? It's been super interesting. It would range from everything from the Super Bowl. For the TV providers, there's this whole freeze around certain times of the year where it takes a

C-level executive to say, Yes, we're going to make this change." But around Superbowl, they're not making any changes. Around Thanksgiving, they're not making any changes. It's important for the data engineering team to think about what's my high watermark, and make sure that your scale to that level, and not just bare minimum, like 1% above that. We're talking about 10%, 20% above that.

Patrick McFadin: Well, it's classic engineering. You don't build a bridge that's exactly to specs. You always make it a 10X fudge factor, right? Because you don't know who's going to drive across it with a big truck. This is not new thinking, and especially with scale and streaming, batch analytics is fine. You can fudge that. When it comes to real time, it's here now, deal with it now, or you missed it.

Jesse Anderson: The better way to think about this is, this is a problem I don't want to solve. You know what? I like to ask people, how many times have you seen on a corporate filing saying, "And I'd like to throw some props out to the data engineers for standing up Cassandra and for writing the distributed system."

Patrick McFadin: Said no one ever.

Jesse Anderson: Said no one ever. It's because it's an infrastructure thing. It's, the value for our business is to use these systems not to create your own system, and that's where you should be. You should be left standing on top of the Cassandra, standing on top of these Flinks and CAFCA and Pulsars, because using them is where you get the ability to scale. It's not writing your own, it's not trying to do some back flips on that. Use it for what it was intended for, and then you don't have to specifically worry about these Black Fridays. Did you do everything? Did you do it right? More than likely you can scale, test it out, test it, stand on the back of those elephants. Stand on top of a person with their PhD, or their 20 years of experience in distributed systems. Don't have your person with their master's degree fresh out of college, just doing this themselves.

Patrick McFadin: Knowledge versus skills, right?

Jesse Anderson: Yup.

Patrick McFadin: All right, so we can put a quick plug here for your book. I think it's great that you're doing this book. I think it's much needed in this industry. So, this is O'Reilly.

Jesse Anderson: It's on O'Reilly, yes.

Patrick McFadin: When is it published? Oh, do I sound like your editor now?

Jesse Anderson: No, so there's a longer story that we'll talk about offline.

Patrick McFadin: Offline, offline, offline.

Jesse Anderson: But yeah, I'm hoping it's sometime this summer.

Patrick McFadin: Okay. Is it available in the early cut or whatever? No, not yet?

Jesse Anderson: Not yet.

Patrick McFadin: Oh, we got stories to tell. All right. Hey, when you're going on summer vacation and you need a good book to read, Jesse's book on data teams would be the right choice.

Jesse Anderson: It will be.

Patrick McFadin: It will be an awesome choice.

Jesse Anderson: It will be the next best seller. It'll be like a Dan Brown thriller.

Patrick McFadin: Oh wow. Okay. Yeah.

Jesse Anderson: I won't oversell it, but it'll probably be better than Da Vinci Code, or maybe even the Vince Flynn.

Patrick McFadin: You know that's not overselling it. That's just being humbly honest. I appreciate that.

Jesse Anderson: I'm by far, the most humble person in the room.

Patrick McFadin: And you know, you pass out $10 billion like he got it. What a player.

Jesse Anderson: I did make it rain rain with $10 billion.

Patrick McFadin: You made it rain with $10 billion. I appreciate that. All right. Thanks Jesse. I really appreciate you being on today.

Jesse Anderson: No, thank you.